UNIVERSITY OF CALIFORNIA,
IRVINE

A Bouquet of Essays

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Economics

by

Michael Ryan Guggisberg

Dissertation Committee:
Professor Dale Poirier, Chair
Professor David Brownstone
Professor Ivan Jeliazkov

2017

DEDICATION

Dedicated to Neko. Despite constantly waking me up in the middle of night, vomiting on my shoes and pooping outside the litter box, I could not ask for a better cat.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

This dissertation would not have been possible without the endless support from my family, friends and advisors. From the beginning of my first year in graduate school, my dissertation chair, Dale Poirier, has been a constant source of guidance and inspiration. He has spent considerable time and effort helping me become the econometrician I am today. Dale's ability to be aware of the state-of-the-art advances in econometrics has shown me the value of staying informed about how the field develops and adapts. I will be forever indebted to him.

My other committee members were also invaluable sources of guidance for me. They have been role models with traits worth striving to achieve. David Brownstone was one of the first professors I worked with. He was always willing to help me tackle any problem I might be having. He would provide a unique, fresh and pragmatic insights that I never would have considered on my own. He has a can do attitude and would never shoot down an idea without providing a helpful alternative. Ivan Jeliazkov taught me the importance of being both a sociable human being and a respectable econometrician at the same time. He also taught me the importance of having a well-rounded skill set.

There have been numerous people outside my committee that have been essential for my success in graduate school. My undergraduate advisor and now friend, Kevin Reffett, encouraged me to go to graduate school. He has since been a constant source of support and encouragement throughout my entire time at UCI. His invaluable personal and professional advice has changed my life forever for the better. He has a terrific taste of music and is an expert in enjoying the pleasures of life. I have tremendous appreciation and respect for him. Daniel Gillen taught me the importance of being strong in both empirical and theoretic econometrics. He inspired me with the idea that resulted in the theoretic findings in my discrete choice paper. Linda Cohen was always happy to meet with me and was very patient with helping me understand the nuances of the issues and processes of the United States Supreme Court. She was instrumental from start to finish in helping me develop that paper. Damon Clark pointed me towards the Project STAR dataset for the application in my third chapter and provided me with helpful advice on how to present academic papers. Matthew Harding also helped me with my academic presentations.

I have many friends who have provided a solid foundation of support for me. I wish I could write something about each one of them but this acknowledgments section has to end at some point. I would like to thank Tim Duffy, Tyler Boston, Ian Burn, Fulya Ozcan, Amine Mahmassani, Sarah Cross, Vanessa Wall, Steven Brownlee (the best officemate), Jason Ralston, Tim Wong, Nanneh Chehras, Jennifer Muz, Patrick Button, William Denetclaw, Andrew Wong, Julie Baker and Elyssa Haeussler. Malorie Hughes has been my greatest friend and source of encouragement. I am incredibly lucky to have met her. I owe my work ethic to the many Marines who helped develop me over the years. Semper Fidelis.

My parents and siblings have provided constant love and support. They have been there for me through thick and thin and helped me see the good side of everything. Their happy and

cheerful attitude to everything has brought light to some of the darkest moments in my life. I love them dearly. And of course, I would like to thank my cat, Neko. She has been there for me since the beginning...of when I got her.

After writing this I am astounded at how many people have had a pivotal impact on my life. I am a product of everyone in this section and am proud to have met each and every one of them.

Additionally, I would like to thank Jason Abrevaya editor of the Journal of Econometric Methods for giving me permission to include the discrete choice paper in my dissertation. I would also like to thank the UCI School of Social Sciences for their funding support and fellowships.

CURRICULUM VITAE

Michael Ryan Guggisberg

EDUCATION

Doctor of Philosophy in Economics **2017**
University of California, Irvine *Irvine, California*

Master of Arts in Economics **2015**
University of California, Irvine *Irvine, California*

Master of Science in Statistics **2015**
University of California, Irvine *Irvine, California*

Bachelor of Science in Economics **2012**
Arizona State University *Tempe, Arizona*

TEACHING EXPERIENCE

Teaching Assistant **2012–2017**
University of California, Irvine *Irvine, California*

ABSTRACT OF THE DISSERTATION

A Bouquet of Essays

By

Michael Ryan Guggisberg

Doctor of Philosophy in Economics

University of California, Irvine, 2017

Professor Dale Poirier, Chair

This bouquet of essays contains four chapters.

In the first chapter I present a brief summary of the literature of misspecified models. I discuss what various estimators are actually estimating when the model is misspecified. Further I discuss corrections to standard errors and when they are useful. I briefly cover hypothesis testing in the presence of misspecified models. I cover both frequentist and Bayesian approached. I show that a misspecified model can indeed be useful and discuss some misconceptions with misspecified models.

The second chapter investigates the impact of misspecification in discrete choice models. I derive necessary and sufficient conditions for consistency of the maximum likelihood estimator from the misspecified model. A corollary is that the misspecified estimator is consistent for the correct sign, under certain conditions. It also follows that Huber-White standard errors can be used to obtain asymptotically conservative type one errors when testing the nullity of the coefficient.

The third chapter builds a Bayesian model for multiple-output quantiles using a commonly accepted definition for the quantile. The prior can be elicited as the ex-ante knowledge of Tukey depth, the first prior of its kind. I apply the model to the Tennessee Project STAR

experiment and find there is a joint increase in *all quantile subpopulations* for reading and mathematics scores given a decrease in the number of students per teacher. This result is consistent with, and much stronger than, the results from previous studies.

The four chapter I investigate if United States Supreme Court Justices recuse themselves strategically. I create a new structural model of recusals. Using this model I find causal evidence that justices recuse themselves strategically. I then calibrate and simulate the model to find the frequency of cases where at least one justice has a conflict of interest but does not recuse. I find at most 47% of cases have at least one justice with a conflict of interest that did not recuse.

It was difficult to come up with an overarching theme for this bouquet of essays – hence the title. The closest theme would be 'model misspecification.' The first chapter provides an overview of misspecified models, the second chapter investigates a misspecified discrete choice models and the third chapter purposefully uses a misspecified model to get an interesting estimator. However, the closest the fourth chapter gets to 'model misspecification' is the use of the Kullback-Leibler distance in a simulation. The Kullback-Leibler distance is often used in the misspecified literature but there is nothing about the distance that makes it inherently related to misspecified models.

Chapter 1

A Brief Overview of Misspecified Models

1.1 Introduction

In this chapter I provide a brief overview of the study of misspecified models within econometrics. Econometrics is an inherently imprecise science. No one expects an econometric model to be an entirely accurate representation of reality. This leads us to question if we can still obtain useful inferences from our inherently flawed models. The answer to this question (as almost all questions in economics) is – it depends. If we are careful about our modeling process we can still derive economically meaningful inferences using incorrect models.

The topic of model misspecification is a very large one, thus this review makes no attempt to be fully comprehensive. I focus attention mostly to likelihood based misspecification and will skip discussion of missing data, measurement error and most nonparameterics. Technical details will be omitted, but are available within the cited papers. Another survey similar to this is one is Monfort (1996). Through 10 examples he shows how misspecified models can

play a role in statistical analysis.

1.2 Preliminaries

Suppose a random variable $Y \in \Re^N$ is generated from the distribution $F^0(y)$ with density $f^0(y)$, which may or may not condition on covariates $X \in \Re^{n \times k}$. This is sometimes called a Data Generating Process or DGP. A researcher chooses a probability model $F(y|\theta)$ with density $f(y|\theta)$ parameterized by a possibly infinite dimensional $\theta \in \Theta$. The model F is correctly specified if there exists a θ^0 such that $f^0(y) = f(y|\theta^0)$ almost everywhere. A model is considered misspecified if there does not exist such a θ^0. The next two examples illustrate this definition.

Suppose $N = 1$ and Y is generated from $Exp(1)$ and the researcher models Y with $Exp(\theta)$. This model is correctly specified because $F^0(y) = 1 - e^{-y}$ equals $F(y|\theta) = 1 - e^{-\theta y}$ for all y when $\theta = \theta^0 = 1$. Suppose instead the researcher modeled Y with $N(\theta, 1)$. This model is misspecified because the support of Y is $[0, \infty)$ but the model $N(\theta, 1)$ produces a positive probability for negative values of Y for any $\theta \in \Theta$.

Define $1_{(A)}$ to be 1 if A is true and 0 if false. One key point in the definition of a misspecified model is that equivalence is almost everywhere. For example, suppose $N = 1$ and Y is generated from a $Unif(0, 1)$ distribution (i.e., $f^0(y) = 1_{(y \in [0,1])}$) but we model Y with the density $f(y|\theta) = 1_{(y \in [0,1] \text{ and irrational})}$. Then f is a correctly specified model because the set of points where $f^0(y) \neq f(y|\theta)$ is the rational numbers which have Lebesgue measure 0. This shows there can be more than one correctly specified model for any DGP.

1.3 History

The first discussion of statistical model misspecification came from economics by Theil (1957). Recognizing that any economic hypothesis is wrong in some way, he explored the effects of various misspecifications in the linear model estimated by OLS. He showed that omitting a single relevant variable can produce bias in all the remaining coefficient estimators and also explored what can happen when one fails to specify a quadratic term or a logarithmic transformation. He also discussed the conditions required to be able to correctly estimate a reduced form elasticity of substitution without using structural demand equations.

Griliches (1957) took Theil's framework and applied it to Cobb-Douglas production functions. He provided conditions for when omission of a relevant variable will produce a positive or negative bias. From his derivations he recommended that one aggregates microvariables with geometric means rather than arithmetic means because it reduces potential bias. Unknowingly to him, this was the first discussion of model robustness.

The next step forward was taken by Rao (1971) who investigated the effects of omitting relevant variables or including irrelevant variables on the standard error and Mean Square Error (MSE) of remaining included estimators. He found that omitting a relevant regressor decreases the standard error of other OLS estimators. Further the MSE is also decreased provided the value of the omitted parameter is smaller than its standard error if it were estimated. Inclusion of an irrelevant variable does not introduce bias but it does increase the standard error and MSE of other included estimators. These results require the regressors to be fixed which is generally not the case in economics. The appendix presents results allowing for stochastic regressors. Deegan Jr. (1976) expands on Rao's work by investigating the bias and MSE when irrelevant variables are included and relevant variables are excluded simultaneously.

The discussion of standard errors for misspecified models typically starts with Eicker et al. (1963) who showed, with fixed regressors, that the least squares estimator is consistent with uncorrelated errors and is asymptotically normal with independent errors. He also provided a consistent estimator of the standard error for the coefficient estimators (i.e., the diagonal of the full covariance matrix). These conditions are much weaker than the traditional Gauss-Markov conditions that require independent and identically distributed errors. This allows the researcher to be fairly agnostic about the distribution of the unobservables. Eicker (1967) extended this result to obtain the full asymptotic covariance matrix of the estimators as well as allowing certain types of serial correlation in the errors. White (1980a), considering only heteroskedasticity, extended Eicker's results to allow for stochastic regressors and derived the asymptotic distribution for arbitrary linear combinations of the estimators. These results do not technically belong in the misspecified model literature, as they are sets of weaker conditions where least squares estimation can still result in correct inferences. However it was the impetus that started the modern research in misspecified models.

The first rigorous treatment of model misspecification for likelihoods was Bayesian. Berk (1966a,b) showed that when the assumed probability model does not contain the DGP, the posterior concentrates (as the number of observations tends to infinity) on a set containing the parameter(s)

$$\theta^\star = \underset{\theta \in \Theta}{argmax}\ E[log(f(Y|\theta))].$$

It is important to note that in the Bayesian context Θ is defined as the parameter values with positive prior density. Berk (1970) provides a weaker set of assumptions for the same result and provides deeper understanding for when the limiting posterior is degenerate at a unique $\theta^\star$. The parameter $\theta^\star$ has an information theoretic interpretation as the parameter minimizing Kullback-Liebler (KL) divergence of the model from the DGP (Kullback and Leibler, 1951; Akaike, 1998). That is

$$\theta^{\star} = \underset{\theta \in \Theta}{argmin}\ E\left[log\left(\frac{f^0(Y)}{f(Y|\theta)}\right)\right].$$

The KL minimizing interpretation is the common interpretation used in modern discussions of misspecified models. The minimized KL divergence is zero if and only if the model is correctly specified, in that case $\theta^{\star} = \theta^0$. Modern treatments of Bayesian misspecified models are covered by Bunke and Milhaud (1998); Kleijn and van der Vaart (2006); Shalizi et al. (2009); Lee and MacEachern (2011); Kleijn and Van der Vaart (2012); Hong and Preston (2012); De Blasi and Walker (2013); Walker (2013); Hoff and Wakefield (2013); Müller (2013); Lv and Liu (2014); Ramamoorthi et al. (2015) and Watson et al. (2016). See Ghosal (1997) for a well written non-technical review of Bayesian asymptotics in the correctly specified case. Chernozhukov and Hong (2003) provide a Baysian framework for M-estimation, which essentially relies on using potentially misspecified likelihoods.

A related topic is the effect of prior specification on the posterior. In the well specified case (i.e., $\theta^0 \in \Theta$) consistency theorems require restrictive conditions on the prior (Doob, 1949; Schwartz, 1965). A common necessary condition is that open Kullback-Leibler neighborhoods of θ^0 must have positive prior support. However this in general is not sufficient and more conditions are required. These conditions on the prior are not imposed just for mathematical convenience. If they are violated then the researcher generally cannot learn about θ^0. This leads us to question for what class of priors can we come to a consensus on θ^0 in the well specified case? When the support of Y is discrete and finite it is necessary and sufficient for the prior to give positive mass to θ^0. If Y is discrete and infinite, continuous or if Θ has infinite dimensional components then more structure is required of the prior to learn θ^0. See Berk (1966b); Freedman (1963, 1965); Freedman and Diaconis (1983); Diaconis and Freedman (1986a,b) for examples of Bayesian inconsistency due to poorly specified priors.

The first rigorous frequentist treatment of model misspecificaiton for maximum likelihood was done by Huber (1967). He showed that under regularity conditions the Quasi-Maximum Likelihood Estimator (QMLE) converges to $\theta^\star$. This parameter is in general not unique. Newey (1987) and Newey and Steigerwald (1997) provide conditions for uniqueness of $\theta^\star$. In most situations the Bayes estimator and the QMLE are asymptotically equivalent. However, there are a few examples where they are different even in parametric models with seemingly well specified priors (Bunke and Milhaud, 1998). Sufficient conditions for the equivalence of the Bayes estimator to equal the QMLE are found in Bunke and Milhaud (1998) and Kleijn and Van der Vaart (2012). Huber's second big finding was deriving the asymptotic sampling distribution of the QMLE. The QMLE is asymptotically Normal with covariance matrix

$$\lim_{n\to\infty} Var(\sqrt{n}\hat{\theta}) = C(\theta^\star).$$

Where $C(\theta) = A(\theta)^{-1}B(\theta)A(\theta)^{-1}$, $A(\theta) = E[\frac{d^2 log(f(Y|\theta))}{d\theta d\theta'}]$, and $B(\theta) = E[\frac{dlog(f(Y|\theta))}{d\theta}\frac{dlog(f(Y|\theta))}{d\theta'}]$. The asymptotic covariance matrix is referred to as the 'sandwich' or 'robust' estimator and is consistently estimated with its empirical counterpart. The importance of this result is that correct standard errors can be obtained using the misspecified model alone. White (1982) derived the same results as Huber under more restrictive, but easier to verify, conditions. However, the covariance matrix he derived was inconsistent and conservative when observations are not IID (Chow, 1984; White, 1983). However, White's standard errors are still used in practice with small sample corrections (MacKinnon and White, 1985; Long and Ervin, 2000). Under a misspecified likelihood, the Bayesian can substitute the posterior covariance with White's covariance matrix to obtain posteriors with smaller asymptotic frequentist risk (Hoff and Wakefield, 2013; Müller, 2013). Newey and West (1987) expanded White's results to allow for heteroskedastic and serially corrected observations and was further refined by Andrews (1991). Freedman (2006) calls into question the usefulness of using robust standard errors. If the model is only slightly misspecified then the robust standard

error is approximately equal to the hessian based standard error. However, if the model is severely misspecified the parameters could be uninteresting and uninterpretable. This will be discussed more in the next section.

There has been much discussion on what the OLS estimator is estimating when the true functional relationship is nonlinear. There is a common misconception that it is a consistent estimator for the first order trend in a Taylor expansion about the mean. This is incorrect. White (1980b) showed that the conditions for consistent estimation of Taylor coefficients requires strong orthogonality and moment restrictions not satisfied in most applications. It is even difficult to obtain the correct information on the sign of the Taylor coefficients with least squares. However, White does find the ordinary least squares estimator is the best linear approximation of the non-linear conditional mean function in mean square error. However, this is only useful for predictive inferences and there is little information contained in the estimated coefficients themselves.

White goes on to show the standard error estimate of the misspecified OLS estimator consistently estimates the sum of the approximation error plus the variance of the independent and identically distributed stochastic errors. Using this result and a derived asymptotic distribution for the OLS estimator, he constructs a test of misspecification. The contribution of this test is that it only relies on the misspecified model. The researcher does not need to consider how the model might be misspecified (conditional on some regularity conditions). Similar tests are proposed in Ramsey (1969); Ramsey and Schmidt (1976); Hausman (1978); White (1980a) and White (1982). Previously common tests for misspecification required the researcher to parametrically specify how the model might be misspecified and then test those specifications directly. However, the drawback of such a broad test such as White's is that it provides little information as to how the model might be misspecified. The second major drawback of this test is it requires the researcher to choose observational weights and little guidance is given on how to choose the weights.

White (1981) extended these results to the non-linear least squares case. He finds the misspecicified non-linear least squares estimator, under regularity conditions, consistently estimates the parameter vector that minimizes mean square error of the true and misspecified mean functions. Both of these results are special cases of White (1982) if the assumed distribution is normal. Domowitz and White (1982) extends these results to dependent observations.

It follows from Huber (1967) and White (1982) that the 'robust' Wald and score (i.e., Lagrange multiplier) tests are asymptotically chi squared and consistent for the KL minimizer. The likelihood ratio test instead is asymptotically distributed as a linear combination of independent chi square distributions under the null hypothesis of the KL minimizer (Foutz and Srivastava, 1977, 1978). Choi and Kiefer (2011) discuss the geometry of the likelihood ratio statistic under misspecification. The conclusions for the Bayesian are similar to the Wald and score tests. Bayesian hypothesis tests and model averages favor the model that minimizes KL divergence (Fernández-Villaverde and Rubio-Ramírez, 2004; Hong and Preston, 2012)

1.4 Interpretation

When the model is misspecified the researcher is estimating a parameter within the parameter space of the assumed model that minimizes KL divergence of the assumed model from the DGP. Since estimates are interpreted with respect to the assumed (incorrect) model, this parameter is in general not interpretable if the model is wrong. However, if the model is only slightly wrong (in a KL sense) then interpretation with respect to the incorrect model may be approximately correct. Freedman (2006) critiques the widespread use of the robust standard error and argues researchers should be focusing more on the estimand (i.e., $\theta^\star$). He states "It remains unclear why applied workers should care about the variance of an estimator for the

wrong parameter." However, there are situations when the assumed incorrect model could be very different (in a KL sense) than the DGP but parameters are still interpretable.

1.4.1 Likelihood

If F is misspecified then we cannot find θ^0 to recover correct inferences on the DGP. However, there can still be parameters representing properties of the DGP that are of interest, call these $\theta^\dagger$. For example, $\theta^\dagger$ could be an expectation of Y or the parameters from a regression of Y on covariates. The next example illustrates this.

Example 1.1. *Suppose $Y \sim F^0$ with positive support, finite first moment and is observed with a random sample of N observations. Let $\theta^\dagger = E[Y]^{-1}$ be the parameter of interest. Let the assumed model be $Exp(\theta)$. Then the KL minimizer $\theta^\star = \underset{\theta}{argmin} - \int ... \int log(\prod_{i=1}^{N} \theta e^{-\theta y_i}) f^0(y_1, ..., y_N) dy_1 .... dy_N$. This is a regular model so the minimum is achieved by setting the derivative to zero and passing the derivative through the integrals. This implies $\frac{d}{d\theta} - log(\theta) + \theta E[Y]\big|_{\theta=\theta^\star} = 0$ and thus $\theta^\star = E[Y]^{-1} = \theta^\dagger$. Therefore the QML of the assumed (potentially misspecified) exponential distribution can consistently estimate the inverse mean of any distribution with positive support and finite first moment.*

In the previous example the QMLE of the exponential distribution consistently estimated the inverse mean of the DGP. Thus the inverse of the QMLE consistently estimates the mean of the DGP (by Mann-Wald). However, one must be careful with interpretation. The parameter of the correctly specified exponential distribution is interpreted as a rate parameter for the rate of arrivals in a Poisson process. If events are not generated from a Poisson process and the implied DGP is not exponential then the QMLE cannot (and should not) be interpreted as a rate parameter. It can only be interpreted as a consistent estimator for the inverse mean of the DGP. For example, if we have a random sample of person heights then the QMLE of the exponential distribution will consistently estimate the inverse mean

height, but interpreting the estimate in terms of rate of arrival is nonsense. Thus if there is a $\theta^\dagger$ of interest and $\theta^\star = \theta^\dagger$ then $\hat{\theta}$ is consistent for these parameter of interest, but one must be careful with interpretation. If $\theta^\star \neq \theta^\dagger$ then usefulness of $\hat{\theta}$ is unclear.

1.4.2 OLS with nonlinear mean

When the assumed model is linear, practitioners sometimes interpret the OLS estimates as first order Taylor approximations of the true conditional mean, see White (1980b) for examples in the literature. This is incorrect unless the true mean function is concave and regressors are fixed, orthogonal, symmetric and have small support – an untenable assumption for a practicing econometrician. However, the OLS estimator can be interpreted as the best linear approximation to the conditional mean with squared error loss. This is a meaningful interpretation of the conditional mean and is useful for interpreting predictions in the face of misspecification. This result was first discovered by Myers and Lahoda (1975) and was further developed and refined by White (1980b) and Bera (1984). However, they provide little discussion on the interpretation of marginal effects. In light of these findings and Freedman's critique Buja et al. (2016b, Section 10) derive an interpretation for OLS coefficients (e.g., marginal effects) in the presence of a non-linear true mean. They show the OLS estimate depends largely on the covariate distribution and observed values of the covariates. For example, two researchers performing the same experiment with the same linear model can arrive at vastly different conclusions in the presence of a true nonlinear mean. The discrepancy is greater than one would obtain from usual statistical variation in the well specified case. Buja and coauthors conclude the model robust interpretation for the OLS coefficient on the first regressor is "Adjusted for all other regressors, the mean deviation of Y in relation to the mean deviation of X_1 is estimated to average between $\hat{\beta}_1$ per 1 unit of X_1." However, the interpretation is obscured by the fact that the averages are weighted averages and the weights depend on the distribution of the covariates. Buja et al. (2016a)

provides generalization to more general types of regression within the IID framework.

1.4.3 Omitted Variable Bias

One common form of misspecification taught in an introductory econometrics class is omitted variable bias. That is, it is the 'bias' induced on the OLS estimator when relevent regressors are omitted. However, the use of the word 'bias' is a misnomer. There is no bias in the true sense of the word. If the researcher has two models $E[Y|X_1, X_2] = \beta_1 X_1 + \beta_2 X_2$ where $\beta_2 \neq 0$ and $E[Y|X_1] = \delta_1 X_1$ and is unsure which one he wants to use he only needs to ask himself what regressors he wants inferences to be conditioned on. If he wishes for his inferences to condition on X_1 and X_2 then the first would be the correct model, if he just wants to condition on X_1 then the second would be the correct model. The two are related by iterated expectations. Let (X_1, X_2) be joint normal random variables with $E(X_1) = E(X_2) = 0$, $Var(X_1) = \sigma_1^2$, $Var(X_2) = \sigma_2^2$ and $Corr(X_1, X_2) = \rho$. Then $E[Y|X_1] = E[E[Y|X_1, X_2]|X_1] = E[\beta_1 X_1 + \beta_2 X_2|X_1] = \beta_1 X_1 + \beta_2 E[X_2|X_1] = \beta_1 X_1 + \beta_2(\frac{\sigma_2}{\sigma_1}\rho X_1) = (\beta_1 + \beta_2\frac{\sigma_2}{\sigma_1}\rho)X_1 = \delta_1 X_1$. The term $\beta_2\frac{\sigma_2}{\sigma_1}\rho$ is taught as the bias from omitting X_2. But this is incorrect, $(\beta_1 + \beta_2\frac{\sigma_2}{\sigma_1}\rho)X_1 = \delta_1 X_1$ is exactly what the mean is supposed to be when not conditioning on X_2, and $(\beta_1 + \beta_2\frac{\sigma_2}{\sigma_1}\rho) = \delta_1$ is exactly the marginal effect of X_1 without controlling for X_2.[1]

When would someone want to condition on X_1 and X_2 or just X_1? This is a common scenario when trying to do causal modeling with observational data. For example, if one were looking to investigate if there is sexual discrimination at a firm, the outcome, Y could be wages, X_1 a female binary variable and X_2 a managerial binary variable, indicating if the individual is a manager. One would anticipate the β_1 parameter to capture the effect of discrimination. The managerial variable would be causally related to wages but should

[1] It would only be bias if one were interpreting their results as if they conditioned on X_1 and X_2. Meaning, they are trying to interpret β_1 and not δ_1.

not be included. Under the hypothesis of sexual discrimination β_2 would capture some of the effect the discrimination since women would be less likely to be hired or promoted to a managerial position. Thus one would not want to include X_2 in the model and would only want to interpret marginal effects of X_1 unconditional on X_2.

1.4.4 Estimand focus

Another approach is to focus interpretation on the estimand. This is common practice in applied econometrics where only moment conditions are required for a model. For example if one were interested in the conditional mean $E[Y|X] = g(x)$ it is sufficient to assume normality of $Y|X$ to consistently estimate the mean via maximum likelihood or Bayesian methods. This holds even if the DGP is non-normal. In fact, this property is shared for all distributions in the exponential family (Gourieroux et al., 1984b). See Gourieroux et al. (1984a) for an application to the Poisson distribution. In another example, the quantile function can be estimated by minimizing the risk function $E[\rho_\tau(Y - \theta)]$ where $\rho_\tau(x) = x(1 - 1_{(x \leq 0)})$. This minimizer is equivalent to the maximizer of the likelihood from an asymmetric Laplace distribution. Thus one can consistently estimate the quantiles of $Y|X$ by assuming $Y|X$ is distributed asymmetric Laplace whether or not the DGP is asymmetric Laplace (Yu and Moyeed, 2001; Yu and Zhang, 2005; Sriram et al., 2013). Yang et al. (2016) provide an asymptotic correction to the variances of the Bayes estimator in the face of misspecification.

1.5 Conclusion

This chapter provided a brief non-technical overview of the literature for misspecified models. In most cases the estimator converges to a value that minimizes KL divergence. If this

parameter is meaningful then consistent standard errors can be obtained using a robust standard error estimator. If the researcher does not know if his model is misspecified he need not throw up his hands and give up. If he is careful with his modeling approach and interpretations then useful inferences can still be obtained in the face of misspecification.

Chapter 2

Misspecified Discrete Choice Models and Huber-White Standard Errors

2.1 Introduction

There are many tools available to statisticians and econometricians to provide reliable inferences for different models. However, these tools come with many nuances and assumptions that can be easily missed or forgotten. One such tool is the Huber-White standard error correction. The Huber-White correction, under certain conditions, can correct a misspecified variance of an asymptotically unbiased maximum likelihood estimator (Huber, 1967; White, 1982). Thus Huber-White standard errors provide asymptotically correct Wald confidence intervals and hypothesis tests. The maximum likelihood estimator of the misspecified conditional logit is not necessarily asymptotically unbiased. The usefulness of a correction for standard errors around an asymptotically biased estimator is unclear (Freedman, 2006). Yet, the Huber-White correction is recommended for misspecified discrete choice models (Train, 2009, pg. 201) and used in practice, see Gartner and Segura (2000); Gould et al. (2004); Ja-

cobs and Carmichael (2002); Lassen (2005). The primary goal of this paper is to investigate the efficacy of the Huber-White correction in discrete choice model and provide conditions when the maximum likelihood estimator of the misspecified model is of interest.

When performing maximum likelihood estimation on a misspecified model (called Quasi-Maximum Likelihood estimation or QML estimation) the estimator is consistently estimating the parameter values that minimizes the Kullback-Leibler (KL) divergence of the misspecified model from the correct model (Huber, 1967; Kullback and Leibler, 1951; White, 1982).[1] The parameter space of the KL minimizer is equivalent to the parameter space of the assumed misspecified model, not the Data Generating Process (DGP). Hence interpretation is reliant on the assumed misspecified model and not the DGP. Thus the KL minimizing values are not necessarily of interest unless there is some econometric information to say otherwise.[2] If the parameter value minimizing KL divergence is not of interest then it is hard to justify the use of the Huber-White correction (Freedman, 2006).[3]

This paper confirms that in general KL minimizer value in misspecified discrete choice models is not a value of interest under most forms of misspecification. I provide necessary and sufficient conditions for when the KL minimizer value is equivalent to parameter value from the DGP.[4] Using this result, I find the misspecified conditional logit consistently estimates the sign of the data generating parameter. It follows that asymptotically correct hypothesis tests and confidence intervals for null coefficients can be obtained using a misspecified

[1] Under some regularity conditions, Huber (1967) finds that the QML estimator is asymptotically normal with standard error that is different than standard maximum likelihood asymptotics. White (1982) derives the same results from Huber (1967) under less general (but more easily verified) assumptions. For the special case of linear models see Eicker (1967) and White (1980a).

[2] For example, the QML estimator for linear exponential families can consistently estimate mean functions as long as the mean of DGP exists (Gourieroux et al., 1984b). The QML estimator of the (misspecified) asymmetric laplace distribution can consistently estimate quantile function as long as the DGP is continuous Yu and Moyeed (2001).

[3] Further, if there is no small sample bias the Huber-White correction can still lead to inconsistent standard errors, requiring adjustment (White, 1983; MacKinnon and White, 1985).

[4] Even though the values will be equal, they are still different parameters. The interpretation of the KL minimzer is with respect to the assumed model and not the DGP. Thus one must take care in how strongly they interpret the resulting estimates.

conditional logit with Huber-White standard errors. These results generalize to arbitrary parametric discrete choice models. Further, if the misspecification is 'small' then the misspecified conditional logit choice probabilities have a smaller Mean Square Error (MSE) than the correctly specified model.

Gourieroux et al. (1984b) provide necessary and sufficient conditions for when the conditional mean function from a misspecified model is consistent for the conditional mean of the DGP. It requires the researcher to correctly specify the conditional mean function and estimate it by maximum likelihood using a member from the linear exponential family. Ruud (1983) provides sufficient conditions for consistency up to a non-zero scale parameter of QML estimators for misspecified binary choice models. Yatchew and Griliches (1985); Cramer (2005) find omitting relevant variables results in inconsistent estimators biased toward zero. However, there is little effect on logit predictions (Ramalho and Ramalho, 2010). My findings agree with and strengthen all these previous studies.[5] Dubin and Zeng (1991) provide a parametric model for incorporating heteroskedasticity which can be used to test for the presence of heteroskedasticity.[6] In practice determining the correct form of heteroskedasticity is not an easy process. McFadden and Train (2000) provide a test for detecting misspecification of random parameters.

Section 2.2 presents the conditional logit and its derivation as a random utility model. Section 2.3 presents the theoretical results for the QML estimator of misspecified discrete choice models. Section 2.4 outlines the simulation procedures for verifying the results in section 2.3. Section 2.5 presents the results from the simulation and section 2.6 concludes.

[5]Ruud (1983) and this paper provide context to the results from **?** for a discrete choice setting. Ruud (1983) provides sufficient conditions for consistency whereas I provide necessary and sufficient conditions. In addition, Ruud (1983) state parameters are identified up to a non-zero scaler. Using the same assumptions and generalizing to multinomial choice I strengthen the result for identification up to a positive scaler, thus preserving sign information

[6] A simulation study by Hole (2006) finds the likelihood ratio and Hessian based Wald tests perform best for detecting heteroskedasticity. Davidson and MacKinnon (1984) conduct a similar study and find the Hessian based score has more reliable performance than the outer product of gradient based score or likelihood ratio tests.

2.2 Models

In this section I present the random utility and conditional logit models.

2.2.1 Random utility

The random utility model is the foundation for much of discrete choice analysis. In this model agent n receives utility U_{nj} from alternative j. If the agent makes only one choice, then they choose the utility maximizing alternative, $j^\star$, denoted $j^\star = \underset{j}{\text{argmax}}\ U_{nj}$. The choice of individual n is represented by $y_{nj^\star} = 1$ and $y_{nj} = 0,\ \forall j \neq j^\star$. This leads to the probability agent n chooses alternative j represented as $P_{nj} = Pr(U_{nj} > U_{ni},\ \forall i \neq j)$. Thus choices are distributed multinomial with likelihood

$$L(\beta|X, y) = \prod_{n=1}^{N}\prod_{j=1}^{J} P_{nj}^{y_{nj}} = \prod_{n=1}^{N} P_{nj^\star}$$

and log likelihood $l(\beta|X, y) \equiv log(L(\beta|X, y)) = \sum_{n=1}^{N} log(P_{nj^\star})$. The utility function considered in this study is additively separable, $U_{nj} = V_{nj} + \epsilon_{nj}$. The observable utility is V_{nj} and the unobservable utility is ϵ_{nj}. The observable portion is a function of observable data and unknown fixed or random parameters.

2.2.2 Conditional logit

Using the setup from above, the conditional logit model assumes the unobserved utility is distributed independent extreme value type 1 (i.e., Gumbel) with location parameter of 0 and scale parameter of 1.[7] This produces the moments $E(\epsilon_{nj}) = \gamma$ and $Var(\epsilon_{nj}) = \frac{\pi^2}{6}$,

[7]There are other assumptions that can be used for the distribution of the unobserved utility. A common one is the normal distribution which leads to the probit model. The probit model has some advantages over the logit model such as not having independence of irrelevant alternatives, allowing for individual specific

where γ is Euler's constant.[8] The observable utility is typically assumed to be linear in (fixed) parameters (i.e., $V_{nj} = X_{nj}\beta$). Then the probability agent n chooses alternative j is $P_{nj} = \frac{e^{X_{nj}\beta}}{\sum_{i=1}^{J} e^{X_{ni}\beta}}$. [9]

2.3 Misspecification of the conditional logit

In this section I provide an (open form) solution to the KL minimizer of the conditional logit (Lemma 2.1). This leads to necessary and sufficient (open form) conditions for the KL minimizer of the conditional logit to be equivalent to the data generating parameter (Theorem 2.1). I extend this result for the KL minimizer for any assumed parametric random utility model (Theorem 2.2). Lastly, I provide a commonly satisfied sufficient condition for the KL minimizer to have the same sign as the DGP parameter (Theorem 2.3) and how this result can be used in practice (Corollary 2.1 and 2.2). Proofs are in the appendix.

Let G be the data generating process with choice probabilities P^0_{nj} and F be the assumed model with choice probabilities $P_{nj}(\beta) = P_{nj}$ (e.g., conditional logit). Then the KL divergence of F from G is

$$KL(G||F) = \sum_{y_1 \in Y_1} \cdots \sum_{y_N \in Y_N} log \left[\frac{\prod_{n=1}^{N} \prod_{j=1}^{J} P_{nj}^{0^{y_{nj}}}}{\prod_{n=1}^{N} \prod_{j=1}^{J} P_{nj}^{y_{nj}}} \right] \prod_{n=1}^{N} \prod_{j=1}^{J} P_{nj}^{0^{y_{nj}}}.$$

where y_{nj} is the jth element of y_n, $Y_n = \{[1, 0, \ldots, 0], [0, 1, \ldots, 0], \ldots [0, 0, \ldots, 1]\}$ and each

correlation over time and random taste variation. However, the interpretation of probit parameters are not as clear as logit parameters, which enjoy a log odds interpretation.

[8] Note that these parameter assumptions are not restrictive since location and scale are unidentified (i.e., $Pr(U_1 > U_2) = Pr(a + bU_1 > a + bU_2)$). However, the assumption of the unobserved utility being distributed extreme value 1 can be argued.

[9] The logit probability form can be derived from the independence of irrelevant alternatives axiom (Luce, 1959). See Train (2009) for a derivation of the conditional logit. The conditional logit can be estimated using iterative maximum likelihood techniques. The conditional logit can be estimated by Bayesian procedures as well (Koop and Poirier, 1993).

element in Y_n has length J (note this is the standard basis in $\Re^J$). Due to independent but not identical likelihood this a very large summation with J^N terms. Fortunately, it simplifies nicely.

Lemma 2.1. $KL(G||F) = \sum_{n=1}^{N} \sum_{j=1}^{J} log\left(\frac{P_{nj}^0}{P_{nj}}\right) P_{nj}^0$

Thus the KL minimizing value is[10]

$$\beta^\star = \underset{\beta}{\text{argmin}}\ KL(G||F) = \underset{\beta}{\text{argmin}} \left[-\sum_{n=1}^{N} \sum_{j=1}^{J} log(P_{nj}(\beta)) P_{nj}^0 \right].$$

Theorem 2.1 provides conditions for the KL minimizer to be equivalent to parameters from the DGP (i.e., $\lim_{n\to\infty} \hat{\beta} = \beta^\star = \beta^0$). Some assumptions and definitions need to be presented first. The first assumption is a DGP assumption providing the framework for how choices could have originated. It is a necessary assumption for $\beta^\star = \beta^0$ to have any meaning.

Assumption 2.1. *The choices are generated from a random utility model where the utility that individual n receives from alternative j is denoted by $U_{nj}^0 = X_{nj}\beta^0 + \eta_{nj}$, $\eta_{nj} \sim F_\eta$ for some F_η. This results in choice probability P_{nj}^0, conditional on X_{nj}.*

The distribution F_η is independent over individuals but can be dependent over alternatives. It could include unmeasurable variables, random effects or different assumptions on the unobserved utility. The next assumption makes explicit the assumed model is conditional logit.

Assumption 2.2. *The assumed choice probability that individual n chooses alternative j is of the form $P_{nj} = \frac{e^{X_{nj}\beta}}{\sum_{i=1}^{J} e^{X_{ni}\beta}}$.*

[10]Notice the minimizer does not necessarily converge with infinite sample size. I suspect under some weak conditions (say covariates satisfy Lindeberg's condition) convergence can be guaranteed.

The covariates, X_{nj}, in Assumptions 2.1 and 2.2 must be the same for $\beta^\star = \beta^0$ to have any useful meaning. Define $W(\beta) = \sum_{n=1}^{N}\sum_{j=1}^{J}\left(P_{nj}(\beta) - P_{nj}^0\right)\begin{bmatrix} X_{nj}^{(1)} \\ \vdots \\ X_{nj}^{(p)} \end{bmatrix}$, $p \equiv length(\beta)$ and $X_{nj}^{(k)}$ the kth element of the row vector X_{nj}. Theorem 2.1 is now presented

Theorem 2.1. *Suppose Assumptions 2.1 and 2.2 hold then* $\lim_{n\to\infty}\hat{\beta} = \beta^\star = \beta^0$ *iff* $W(\beta^0) = \boldsymbol{0}$.

This is not an existence result. There might not exist a β^0 or $\beta^\star$ such that $W(\beta^0) = \mathbf{0}$ (e.g., probit data generating process with $\beta^0 \neq 0$). If $W(\beta^0) = \mathbf{0}$ exists, then the above holds.

The result in Theorem 2.1 can be generalized such that P_{nj} is derived from any random utility model with observable utility $V_{nj} = X_{nj}\beta$. This can be done using a result from McFadden and Train (2000) that shows the mixed logit can approximate any random utility model. We will need to restrict ourselves to a convenient class of random utility models that are of the same form as those in Assumption 2.1. Again, this provides meaning to $\beta^\star = \beta^0$.

Assumption 2.3. *The model is assumed to be a random utility model where the utility that individual n receives from alternative j is denoted by* $U_{nj} = X_{nj}\beta + \epsilon_{nj}$ *and* $\epsilon_{nj} \sim F_\epsilon$ *for some* F_ϵ. *This results in choice probability* $P_{nj}(\beta)$.

Assumption 2.3 is a generalization of Assumption 2.2. The more general theorem is now presented.

Theorem 2.2. *Suppose Assumptions 2.1 and 2.1 hold then* $\lim_{n\to\infty}\hat{\beta} = \beta^\star = \beta^0$ *iff* $W(\beta^0) = \boldsymbol{0}$.

Theorems 2.1 and 2.2 are difficult to operationalize because β^0 and P_{nj}^0 are not known. Additional, P_{nj} might not exist in closed form for Theorem 2.2. Fortunately, it is fairly simple to find when the KL minimizer has the same sign as the DGP parameter. It is sufficient for the KL minimizer to have the same sign as the DGP parameter when the assumed model is conditional logit.

The next assumption is a common identification assumption made in random utility models. It is usually implicitly assumed, but I use it explicitly, so I state it explicitly.

Assumption 2.4. *The fixed covariates vary over alternatives (i.e.,* $\hat{Var}_j(X_{nj}) > 0$*).*

If Assumption 2.4 cannot be satisfied (e.g., the covariate is a measure of an individuals income) then one can interact it with alternative specific constants. Finally, we arrive at the result that assuming the model is conditional logit is sufficient to identify the signs of the coefficients in the data generating process.

Theorem 2.3. *Suppose Assumptions 2.1, 2.2 and 2.4 hold then* $sign(\beta^\star) = sign(\beta^0)$ *element by element.*

Therefore QML estimation of the conditional logit is consistent for the correct sign of the fixed parameters in any random utility framework that is generated according to Assumption 2.1. A corollary follows for the special case of the DGP parameters being 0.

Corollary 2.1. *Suppose assumptions 2.1, 2.2 and 2.4 hold. If* $\beta_k^0 = 0$ *then* $\beta_k^\star = 0$ *for any* $k \in \{1, 2, ..., p\}$.

It follows that hypothesis tests about $\beta_k^0 = 0$ can be conducted with the misspecified conditional logit.

Corollary 2.2. *Suppose assumptions 2.1, 2.2 and 2.4 hold then hypothesis tests of the form* $H_0 : \beta_k^0 = 0$ *vs* $H_a : \beta_k^0 \neq 0$, $H_0 : \beta_k^0 \geq 0$ *vs* $H_a : \beta_k^0 < 0$ *and* $H_0 : \beta_k^0 \leq 0$ *vs* $H_a : \beta_k^0 > 0$ *can be consistently performed with* $\hat{\beta}_k$ *for* $\beta_k^\star = \beta_k^0$ *using the Huber-White standard errors. The type one error rate will be asymptotically conservative (i.e.,* $\lim_{n\to\infty} Pr(reject|H_0) \leq \alpha$*).*

Thus if choices are generated according to Assumption 2.1 a researcher can test for non-zero coefficients using the (possibly misspecified) conditional logit and obtain asymptotically

correct type one errors. This test is only justified for the cases presented in the corollary. It is unclear what the asymptotic type one errors are from testing for some arbitrary non-zero constant (e.g., $H_0 : \beta_k^0 \leq c$ vs $H_0 : \beta_k^0 > c$ where $c \neq 0$). I make no claims for the power level of the tests.[11]

2.4 Simulation

In this section I verify several results under three different data generating processes: conditional logit, mixed logit and heteroskedastic logit. First, I show the Huber-White standard error provides asymptotically correct type one error rates for null coefficients in the conditional logit despite misspecification. Then I show the KL minimizer from the assumed conditional logit model is not equal to the data generating parameters under misspecification. Next I show confidence intervals with Huber-White standard errors do not cover the data generating parameter, β^0, with the appropriate coverage probability. However, they do cover the KL minimizer, $\beta^\star$, with the appropriate coverage probability. Lastly, I compare the MSE of the estimated choice probabilities with the correctly specified models. Subsections 5.1 and 2.5.2 present the mixed logit and heteroskedastic logit data generating processes. Subsection 2.5.3 presents the simulation design.

2.4.1 Mixed logit

The first form of misspecification is a failure to specify the random component in the mixed logit model. The utility function for the mixed logit is $U_{nj} = V_{nj} + Z_{nj}b_n + \epsilon_{nj}$ with random parameter vector $b_n \overset{\text{iid}}{\sim} F_b$ and $E[b_n] = 0$, where F_b is the mixing distribution. The conditional

[11] However, I would anticipate the power of detecting the correct sign to be an increasing function of the magnitude in the correct direction.

probability agent n chooses alternative j is $P_{nj}|b_n = \frac{e^{V_{nj}+Z_{nj}b_n}}{\sum_{i=1}^{J} e^{V_{ni}+Z_{ni}b_n}}$.The unconditional choice probability is $P_{nj} = \int \frac{e^{V_{nj}+Z_{nj}b_n}}{\sum_{i=1}^{J} e^{V_{ni}+Z_{ni}b_n}} dF_b$.[12]

2.4.2 Heteroskedastic logit

The second form of misspecification is failing to account for the heteroskedastic parameter in the heteroskedastic logit model. The heteroskedastic logit model introduces heteroskedasticity by allowing the scale parameter of the extreme value type 1 distribution to vary over individuals and alternatives (i.e., $\theta_{nj}\epsilon_{nj} \sim EV1(1, \theta_{nj})$, $\theta_{nj} > 0$). [13] Thus the utility function is $U_{nj} = V_{nj} + \theta_{nj}\epsilon_{nj}$. Heteroskedasticity over individuals, $\theta_{nj} = \theta_n$, can represent differing abilities for individuals to understand the presented alternatives. This type of heteroskedasticity leads to a simple closed form solution, $P_{nj} = \frac{e^{V_{nj}/\theta_n}}{\sum_{i=1}^{J} e^{V_{ni}/\theta_n}}$.[14] The heteroskedasticity is parameterized $\theta_{nj} = \gamma_0 e^{Z_n^T \gamma_1}$, $(\gamma_0, \gamma_1) \in \Re^{++} \times \Re^k$ where Z_n are data and γ's are parameters. [15] When heteroskedasticity is only over individuals then $\gamma_0 = 1$ for identification.[16]

[12]The integral can be evaluated by quadrature, simulation or MCMC methods (Lange, 1999; Train, 2009; Jeliazkov and Lee, 2010). This paper uses Newton-Rhapson based maximum simulated likelihood by halton methods to mimic STATA procedures from the 'mixlogit' command (Haan and Uhlendorff, 2006; Hole, 2007). Since the integral is simulated, the number of simulations needs to scale appropriately with sample size for asymptotic normality to be achieved. I set the number of integral simulations equal to $N^{0.85}$. Estimation can also be performed using hierarchical Bayesian methods (Dumont and Keller, 2015).

[13]Dubin and Zeng (1991) provides such a model for the generalized extreme value family of models.

[14]Dubin and Zeng (1991) mistakenly reports $P_{nj} = \frac{e^{V_{nj}\theta_n}}{\sum_{i=1}^{J} e^{V_{ni}\theta_n}}$. Alternatively, heteroskedasticity can be over individuals and alternatives, θ_{nj}, representing choice fatigue or different channels to view alternatives. If θ_{nj} then the choice probability does not exist in closed form. It can be evaluated by quadrature (Bhat, 1995) or by laplace transform (Dubin and Zeng, 1991).

[15]A common alternative parameterization is $\theta_{nj} = (1 + Z_{nj}^T\gamma)^2$, $\gamma \in \Re^k$.

[16]This is estimated using iterative maximum likelihood techniques.

2.4.3 Simulation design

The form of the observable utility for all DGPs is $V_{nj} = X_{nj}^{(1)}\beta_1^0 + X_{nj}^{(2)}\beta_2^0$, with data generating parameters $(\beta_1^0, \beta_2^0) = (0, 1)$ for determining type one error rates of a null coefficient (Corollary 2) and $(\beta_1^0, \beta_2^0) = (-2, 1)$ for all other simulations. The data is simulated from $X_{nj}^{(1)} \sim \chi_1^2$, representing alternative price, and $X_{nj}^{(2)} \sim Bernoulli(.8)$, representing an alternative specific constant. I simulate N individuals choosing from J alternatives. The values for (N, J) used in this study are presented in Table 2.1. The number of simulations for each (N, J) pair is $1,000$.

Table 2.1: Total individuals and alternatives

(N,J)	Individuals, N		
Alternatives, J	(100, 2)	(500, 2)	(1000, 2)
	(100, 3)	(500, 3)	(1000, 3)
	(100, 5)	(500, 5)	(1000, 5)

Misspecification is introduced in two different forms. The first form is where the choices are generated according to the mixed effects logit, but the random effect is not modeled. The second form is where the data is generated according to the heteroskedastic logit, but the coefficient against the unobservable utility is not modeled. The form of the random effect in the mixed logit is an alternative specific constant where

$$Z_{nj}b_n = \begin{cases} b_n \sim N(0, \sigma^2) & \text{if } j = 1 \\ 0 & \text{if } j \neq 1 \end{cases}$$

This represents a random preference for an alternative that is presented once to every individual. The values for σ^2 are $\sigma^2 \in \{0.5^2, 1^2, 2^2\}$. In the second DGP choices are simulated from a heteroskedastic logit where the heteroskedastic parameter is $\theta_n = e^{\gamma_1 W_n}$. Where W_n is drawn from a discrete uniform with support from -2 to 2 and $\gamma_1 \in \{.5, 1, 1.5\}$.

The efficacy of the Huber-White correction in this study is determined by type one errors

and coverage probabilities of the β^0 and $\beta^\star$ coefficients from the DGP and KL minimization respectively. If the Huber-White correction provides a 'good' coverage probability then the correction could be useful. Coverage probabilities for Wald-based 80% confidence intervals are computed for the misspecified and correctly specified models. Hessian, Huber-White and simulated standard errors are used. The comparison of interest is to see if the 0.80 coverage probability is better targeted with the Hessian or Huber-White standard errors. Simulation based standard errors are included as a check to see if the distribution of the maximum likelihood estimator (of the correctly specified model) is approaching normality.[17] Since the minimizing parameter changes with the data, the data is kept fixed and only the unobserved utility is redrawn in each simulation.

2.5 Results

The results from the simulation study are presented in this section. All tables can be found in the appendix.

2.5.1 Type one error of null parameter

If the data generating parameter is null ($\beta_1^0 = 0$) then the Huber-White standard error should provide an asymptotically conservative type one error rate with the misspecified conditional logit model (Corollary 2.1). Tables B.1 and B.2 show simulated type one errors of $H_0 : \beta_1 = 0$ in conditional logit. Table B.1 is the mixed logit DGP and Table B.2 is the heteroskedastic logit DGP. A type one error rate close to 0.20 shows good performance.

I find Hessian and Huber-White standard errors result in similar inferences and neither appear to dominate the other. Additionally, the correct type one error seems to be achieved

[17] Note this is only a necessary, not a sufficient, condition for normality.

in most environments except when the random effect is large ($\sigma^2 = 2^2$) then the type one errors are erratic.

Thus the misspecified conditional logit model usually provides correct type one errors for null coefficients. However, there does not seem to be any guidance as whether to use the Hessian or Huber-White standard errors since both perform similarly.

2.5.2 KL minimizing parameter

When the model is misspecified the QML estimator is estimating the parameter minimizing the KL divergence of the assumed model from the DGP. Tables B.3 and B.4 show the KL minimum and minimizer from the simulation where the researcher fails to specify a random effect (Table B.3) or a heteroskedastic effect (Table B.4). The KL minimizer is found by directly minimizing the KL divergence using numerical methods. Simulation standard errors of the QML estimates around the KL minimizing parameter are also included.[18]

As σ^2 and γ_1 increase, there is an increase in the KL divergence of the conditional logit (evaluated at the KL minimizer) from the DGP. The KL divergence (evaluated at the minimizer) can be used to measure the degree of misspecification since the divergence becomes 0 as the DGP converges to the conditional logit. As the divergence increases the KL minimizer attenuates toward 0. This complements the results from Cramer (2005); Yatchew and Griliches (1985). Lastly, the standard error of the QML estimator around the KL minimizing parameter decreases. This result is fairly surprising. It is saying a more misspecified model is more informative about the KL minimizer.

As the number of individuals increases (N increases) the distance between the KL minimizer and the DGP parameters fluctuates. This can be explained by the possibility the KL mini-

[18] Calculated by taking the standard deviation over the monte carlo simulation. The confidence intervals based on this standard error assumes normality.

mizer does not necessarily converge with N. As mentioned earlier, I anticipate convergence under some (unspecified) mild conditions.

As the number of alternatives increases (J increases) the distance between the KL minimizer and the DGP parameters tends to decrease. This can be explained by the fact that the misspecification for the random effect is only a misspecification on the first alternative, and hence, the misspecification gets 'washed out' as the number of alternatives increases. Likewise the misspecified heteroskedasticity is a misspecification over individuals and hence becomes 'less misspecified' as the number of alternatives increases.

2.5.3 Coverage probabilities of β^0

Coverage probabilities of the DGP parameters, β^0, are presented in Tables B.5, B.6, and B.7. The tables show the coverage probabilties for 80% confidence intervals using Hessian, Huber-White and simulation based standard errors. Misspecified and correctly specified models are considered.

Coverage probabilities from the correctly specified conditional logit can be seen in Table B.5. The Hessian and Huber-White confidence intervals appear to perform the same and are close to the target coverage probability.

Coverage probabilities from the misspecified conditional logit when the DGP is mixed logit and heteroskedastic logit are presented in Table B.6 and B.7. Coverage probabilities of the correctly specified mixed and heteroskedastic logit models are also included in the tables. There tends to be little difference between the coverage probabilities of Hessian and Huber-White standard errors. As σ^2 and γ_1 increase the coverage probabilities for the misspecified model decrease and the coverage probabilities of the correct models are erratic. For small σ^2 ($\sigma^2 = 0.5^2$) the misspecified model target the 0.80 coverage probability better than the

correctly specified mixed logit.

As the number of individuals increase (N increases) the confidence intervals for the misspecified model cover the DGP parameters less. This is more apparent the larger σ^2 or γ_1 is. The coverage probabilities for the correct mixed logit are erratic and do not show any sign of converging to the targeted value.

A surprising finding was that the misspecified conditional logit performed better than the correctly specified mixed logit for small σ^2. This is likely caused by the increased estimation error induced by simulation of the integral in the likelihood. This also confirms the finding in Keane (1992) and Ruud (1996) that if the random effect is not 'big enough' the model is nearly unidentified (these papers only explored this result for the probit). It was also interesting to note that the coverage probabilities for the parameters estimated by maximum simulated likelihood of the correctly specified model were erratic. This might be fixed by using a different scaling for the number of integral simulations.

While the Huber-White standard errors provided slightly better coverage in the misspecified heteroskedastic model, they did not provide a reliable correction to obtain the target 0.80 coverage probability. The finite sample coverage from the correctly specified heteroskedastic logit model using Huber-White standard errors performed was slightly worse than the Hessian standard errors (but did asymptotically target the 0.80 coverage probability). In general, the coverage probabilities were less affected by failing to specify the random effect than the heteroskedastic effect. This could be since the random effect is on an alternative specific constant for only one presented alternative. A random coefficient against a continuous covariate might produce a larger deviation from the targeted coverage probability.

2.5.4 Coverage probabilities of $\beta^\star$

The results from Huber (1967) and White (1982) tell us that under certain assumptions the QML estimator in the misspecified model is asymptotically normal centered on the KL minimizing parameter with Huber-White covariance. The KL minimizing parameter is not necessarily the DGP parameter value. If the coverage probabilities are calculated for the KL minimizing parameter instead of the DGP parameter then the confidence intervals with Huber-White standard error should target the 0.80 coverage probability. Tables B.8 and B.9 show the coverage probabilities of the KL minimizing parameter in confidence intervals from the misspecified models. The KL minimizing parameter is calculated by using numerical methods to minimize the KL divergence of the misspecified model from the correctly specified model.[19]

Coverage probabilities based on the Huber-White standard error tend to be conservative (greater that 0.80) especially for larger σ^2 and γ_1. This verifies the result in White (1983). The Hessian standard errors tend to lead to anti-conservative coverage probabilities (less than 0.80). As σ^2 and γ_1 increase, the coverage probabilities based on the Huber-White estimator tended to increase and become more conservative. Coverage probabilities based on the Hessian would fluctuate. The simulation based standard errors show that finite sample approximate normality appears to occur for the QML estimator.

The Huber-White standard errors tend to perform slightly better (although conservatively) at capturing the KL minimizing parameters. However, this estimand may not be of interest.

[19] Coverage of the KL minimizer of the correctly specified conditional logit is not included because the model is correctly specified and thus the KL minimizer is equivalent to the DGP parameters.

2.5.5 MSE of choice probabilities

Previous research has found model misspecification has little effect on the estimated choice probabilities Ramalho and Ramalho (2010). To investigate if this is true, I compare the MSE of the choice probabilities from the misspecified conditional logit with the properly specified mixed logit (Table B.10) and heteroskedastic logit (Table B.11).

The MSE of the misspecified conditional logit for the mixed logit DGP is lower than the correct model; this holds for all N, J, σ^2. The conditional logit has a decreasing MSE with increased sample size. The mixed logit has a fluctuating MSE. The MSE of the misspecified conditional logit for the heteroskedastic logit DGP has a smaller MSE than the correctly specified heteroskedastic logit for small heteroskedasticity ($\gamma_1 = 0.5$), but this relationship flips for larger heteroskedasticity ($\gamma_1 \in \{1, 1.5\}$).

The MSE of the misspecified conditional logit decreases with increasing number of alternatives but fluctuates with increasing number of individuals. The MSE of the correctly specified heteroskedastic logit tends to decrease with increase number of individuals or alternatives.

These results suggest that unless the misspecification is 'large enough' the added noise from simulating a random effect or estimating additional parameters adds excessive error. Thus it is better to estimate a a misspecified conditional logit in that case.

2.6 Discussion

If a researcher believes to have misspecified a discrete choice model, the Huber-White correction will not help recover correct inferences on the DGP parameters in general. If the researcher believes they specified the model correctly but wishes to use the Huber-White correction 'just to be safe' then the correction will work just-as-well (at best) as the Hessian

based confidence interval. The KL minimizer appears to diverge from the data generating parameters with increasing misspecification and it is unclear when the KL minimizer will equal the DGP parameter in general. However, if the DGP parameter is zero then the KL minimizer is also zero. Thus the Huber-White correction can be justified for obtaining conservative type one errors when testing for positive, negative or zero coefficients with a conditional logit. If the researcher is interested in estimating choice probabilities then the researcher should use the possibly misspecified conditional logit unless they think that model is excessively 'far' from the DGP.

A possible extension is to find more interpretable conditions when the KL minimizer equals the data generating parameter. Finding bounds on the difference of the DGP parameter and the KL minimizer could provide guidance as to what misspecified parametric discrete choice models would be preferred. Lastly, the effect of misspecification on marginal effects was not explored in this paper. While marginal effects are of secondary interest (compared the sign of the parameter), they do provide rich economic interpretations and it would be interesting to see how they would be affected.

Chapter 3

A Bayesian Approach to Multiple-Output Quantile Regression

3.1 Introduction

Univariate quantile regression was originally proposed by Koenker and Bassett (1978) and has since become a popular mode of inference among empirical researchers (see Yu et al. (2003) for a survey). Additionally, econometricians and statisticians have brought many methodological advances to the field. One such advance was the introduction of quantile regression into a Bayesian framework (Yu and Moyeed, 2001). This advance opened the doors for Bayesian inference and generated a series of applied and methodological research.[1]

The literature on multivariate quantiles has been growing slowly but steadily since the early 1900s (Small, 1990). Much of the reason for the slow growth is because a multivariate quantile can be defined in many different ways and there has been little consensus on which

[1]For example, see Alhamzawi et al. (2012); Benoit et al. (2014); Benoit and Van den Poel (2012); Feng et al. (2015); Kottas and Krnjajić (2009); Kozumi and Kobayashi (2011); Lancaster and Jae Jun (2010); Rahman (2016); Sriram et al. (2013); Taddy and Kottas (2010); Thompson et al. (2010).

is the most appropriate (Serfling, 2002). Further, the literature for Bayesian inference in this field is sparse. Only two papers exist and neither use a commonly accepted definition for a multivariate quantile.[2]

I present a Bayesian framework for multivariate quantiles defined in Hallin et al. (2010). Their 'directional' quantiles have theoretic and computational properties not enjoyed by many other definitions. My approach uses an idea similar to Chernozhukov and Hong (2003), it assumes a likelihood that is not necessarily representative of the Data Generating Process (DGP). However, I show the resulting posterior converges almost surely to the true value.[3] By performing inference in this framework one gains the advantages of a Bayesian analysis. The Bayesian machinery provides a principled way of combining prior knowledge with data to arrive at conclusions. This machinery can be used in a data-rich world, where data is continuously collected, to make inferences and update them in real time.[4]

The prior is a required component Bayesian analysis where the researcher elicits their pre-analysis beliefs for the population parameters. The prior in this model is closely related to the Tukey depth of a distribution (Tukey, 1975). Tukey depth is a notion of multivariate centrality of a data point. This is the first Bayesian prior for Tukey depth. Once a prior is chosen, estimates can be computed using MCMC draws from the posterior. If the researcher is willing to accept prior joint normality of the model parameters then a Gibbs MCMC sampler can be used. Gibbs samplers have many computational advantages over other MCMC algorithms. Consistency of the posterior and a Bernstein-Von Mises result are verified via a small simulation study.

[2] Drovandi and Pettitt (2011) uses a copula approach and Waldmann and Kneib (2015) uses a multivariate Asymmetric Laplace likelihood approach.

[3] Posterior convergence means that as sample size increases all the probability mass for the posterior is concentrated in smaller neighborhoods around the true value. Converging eventually to a point mass at the true value.

[4] Additionally, Bayesians can make exact finite sample inferences, the Bayesian posterior interval has a more intuitive interpretation than a Frequentist confidence interval and full predictive distributions can be obtained using Markov Chain Monte Carlo (MCMC) draws. There is a host of other advantages including computation, hypothesis testing, handling nuisance parameters and introducing hierarchy into a model.

Lastly, the model is applied to the Tennessee Project STAR experiment (Finn and Achilles, 1990). The goal of the experiment was to determine if classroom size has an effect on learning outcomes.[5] I compare the results of class size on test score by estimating the quantiles of mathematics and reading test scores for students in the first grade. I find *all quantile subpopulations* of mathematics and reading scores improve for students in smaller classrooms. This result is consistent with, and much stronger than, the result one would find with multivariate linear regression. An analysis by multivariate linear regression finds mathematics and reading scores improve *on average*, however there could still be subpopulations where the score declines.[6] The multiple-output quantile regression approach confirms there are no quantile subpopulations where the score declines. This is truly a statement of 'no child left behind' opposed to 'no average child left behind.'

3.1.1 Quantiles and quantile regression

Quantiles sort and rank observations to describe how extreme an observation is. In one dimension, for $\tau \in (0,1)$, the τth quantile is the observation that splits the data into two bins: a left bin that contains $\tau \cdot 100\%$ of the total observations that are smaller and a right bin that contains the rest of the $(1-\tau)\cdot 100\%$ total observations that are larger. When expanding to higher dimensions, the notion of partitioning the data into two sets is maintained. The entire family of $\tau \in (0,1)$ quantiles allows one to uniquely characterize the full distribution of interest. A population univariate quantile is defined as follows: let $Y \in \Re$ be a univariate random variable with Cumulative Density Function (CDF), $F_Y(y) = Pr(Y \leq y)$ then the

[5] Students were randomly selected to be in a small or large classroom for four years in their early elementary education. Every year the students were given standardized math and reading tests.

[6] A plausible narrative is a poor performing student in a larger classroom might have more free time due to the teacher being busy with preparing, organization and grading. During this free time the student might read more than they would have in a small classroom and might perform better on the reading test than they would have otherwise.

τth quantile is

$$Q_Y(\tau) = \inf\{y \in \Re : \tau \leq F_Y(y)\}. \tag{3.1}$$

If Y is a continuous random variable then the CDF is invertible and the quantile is $Q_Y(\tau) = F_Y^{-1}(\tau)$. Whether or not Y is continuous, $Q_Y(\tau)$ can be defined as the generalized inverse of $F_Y(y)$ (i.e. $F_Y(Q_Y(\tau)) = \tau$).[7] The definition of sample quantile is the same as (3.1) with $F_Y(y)$ replaced with its empirical counterpart.

Quantiles can be computed via an optimization based approach. This is somewhat surprising because quantiles are a notion of ranking and sorting—a link to optimization is not immediately clear. This relationship between quantiles and optimization was first shown in Fox and Rubin (1964). Define the check function to be

$$\rho_\tau(x) = x(\tau - 1_{(x<0)}), \tag{3.2}$$

where $1_{(A)}$ is an indicator function for event A being true. It can be shown the τth population quantile of $Y \in \Re$ is equivalent to $Q_Y(\tau) = \underset{a}{argmin}\, E[\rho_\tau(Y-a)]$. Note this definition requires $E[Y]$ and $E[Y1_{(Y-a<0)}]$ to be finite. The corresponding sample quantile estimator is

$$\hat{\alpha}_\tau = \underset{a}{argmin} \frac{1}{n} \sum_{i=1}^{n} \rho_\tau(y_i - a). \tag{3.3}$$

If the moments of Y are not finite, an alternative but equivalent definition can be used instead (Paindaveine and Šiman, 2011).

Univariate linear conditional quantile regression (generally known as 'quantile regression') was originally proposed by Koenker and Bassett (1978). They define the τth conditional

[7] There are several different ways to define the generalized inverse of a CDF and each has different properties (Embrechts and Hofert, 2013; Feng et al., 2012).

population quantile function to be

$$Q_{Y|X}(\tau) = \inf\{y \in \Re : \tau \leq F_{Y|X}(y)\} = X'\beta_\tau \tag{3.4}$$

which can be equivalently defined as $Q_{Y|X}(\tau) = \underset{b}{argmin}\, E[\rho_\tau(Y - X'b)|X]$ (provided the moments $E[Y|X]$ and $E[Y1_{(Y-X'b<0)}|X]$ are finite). The parameter β_τ is estimated in the frequentist framework by solving

$$\hat{\beta}_\tau = \underset{b}{argmin}\, \frac{1}{n}\sum_{i=1}^{n} \rho_\tau(y_i - x_i'b). \tag{3.5}$$

This optimization problem can be written as a linear programming problem and solutions can be found using the simplex or interior point algorithms.

There are two common motivations for quantile regression. The first is its estimates and predictions are robust to outliers and certain violations of model assumptions.[8] The second is specific quantiles can be of greater scientific interest than means or conditional means (as one would find in linear regression).[9] These two motivations also apply to multiple-output quantile regression. See Koenker (2005) for a well written survey of the field of univariate quantile regression.

There have been several approaches to generalizing quantiles from a univariate to a multivariate case. This generalization is difficult because the univariate quantile can be defined as a generalized inverse of the CDF. Since a multivariate CDF has multiple inputs and hence, is not one-to-one, then a definition based off inverses can lead to difficulties. See Serfling and Zuo (2010) for a discussion of desirable criteria one might expect a multivariate quantiles to have and Serfling (2002) for a survey of extending quantiles to the multivariate case. Small

[8]For example, the median of a distribution can be consistently estimated whether or not the distribution has a finite first moment.

[9]For example, if one were interested in the effect of police expenditure on crime, one would expect there to be larger effect for high crime areas (large τ) and little to no effect on low crime areas (small τ).

(1990) surveys the special case of a median.

This paper follows a framework of multivariate quantiles using a 'directional quantile' approach introduced by Laine (2001) and rigorously developed by Hallin et al. (2010). A directional quantile of $\mathbf{Y} \in \Re^k$ is a function of two objects: a direction vector $\mathbf{u}$ (a point on the surface of k dimension hypersphere) and a depth $\tau \in (0, 1)$. A directional quantile is then uniquely defined by $\boldsymbol{\tau} = \mathbf{u}\tau$. The $\boldsymbol{\tau}$ directional quantile hyperplane is denoted $\lambda_{\boldsymbol{\tau}}$ which is a hyperplane through $\Re^k$. The hyperplane $\lambda_{\boldsymbol{\tau}}$ generates two quantile regions: a lower region of all points below $\lambda_{\boldsymbol{\tau}}$ and an upper region of all points above $\lambda_{\boldsymbol{\tau}}$. The lower region contains $\tau \cdot 100\%$ of observations and the upper region contains the remaining $(1-\tau) \cdot 100\%$. Additionally, the vector connecting the probability mass centers of the two regions is parallel to $\mathbf{u}$. Thus $\mathbf{u}$ orients the regression and can be thought of as a vertical axis.

3.1.2 Bayesian single-output quantile regression

A Bayesian approach to quantile regression may seem inherently contradictory to Bayesian principles. Bayesian methods require a likelihood and hence a distributional assumption, yet one common motivation for quantile regression is to avoid making distributional assumptions. Yu and Moyeed (2001) introduced a Bayesian approach by using a (possibly misspecified) likelihood of an Asymmetric Laplace Distribution (ALD), whose maximum likelihood estimate is equal to the estimator from (3.5). The Probability Density Function (PDF) of the ALD is

$$f_\tau(y|\mu, \sigma) = \frac{\tau(1-\tau)}{\sigma} exp(-\frac{1}{\sigma}\rho_\tau(y-\mu)). \tag{3.6}$$

A Bayesian assumes $Y|X \sim ALD(X'\beta_\tau, \sigma, \tau)$, selects a prior, and performs estimation using standard procedures. Sriram et al. (2013) showed posterior consistency, meaning as sample size increases the probability mass of the posterior concentrates around the values of β that

satisfy (3.4). Yang et al. (2016) found consistent variances can be achieved using a simple modification to the posterior using the draws from the MCMC algorithm. If one is willing to accept joint normality of β_τ then a Gibbs sampler can be used to obtain random draws from the posterior (Kozumi and Kobayashi, 2011). If regularization is desired, then an adaptive Lasso sampler can be used (Alhamzawi et al., 2012). Nonparametric Bayesian approaches to quantile regression have been proposed by Kottas and Krnjajić (2009) and Taddy and Kottas (2010).

3.2 Multiple-output quantile regression

This section presents the multiple-output quantile regression and discusses some of its properties. An example is presented at the end of this section to aid in the explanation. The rest of the exposition follows closely from Hallin et al. (2010). Let $[Y_1, Y_2, ..., Y_k]' = \mathbf{Y}$ be a k-dimension random vector. The direction and magnitude of the directional quantile is defined by $\boldsymbol{\tau} \in \mathcal{B}^k = \{\mathbf{v} \in \Re^k : 0 < ||\mathbf{v}||_2 < 1\}$. Where $\mathcal{B}^k$ is a k-dimension unit ball centered at $\mathbf{0}$ (with center removed). Define $||\cdot||_2$ to be the l_2 norm. The vector $\boldsymbol{\tau}= \tau\mathbf{u}$ can be broken down into two components: direction, $[u_1, u_2, ..., u_k]' = \mathbf{u} \in \mathcal{S}^{k-1} = \{\mathbf{v} \in \Re^k : ||\mathbf{v}||_2 = 1\}$ and magnitude, $\tau \in (0, 1)$.

Let $\boldsymbol{\Gamma}_\mathbf{u}$ be some $k \times (k-1)$ matrix such that $[\mathbf{u} \vdots \boldsymbol{\Gamma}_\mathbf{u}]$ is an orthonormal basis of $\Re^k$. Further define $\mathbf{Y}_\mathbf{u} = \mathbf{u}'\mathbf{Y}$ and $\mathbf{Y}_\mathbf{u}^\perp = \boldsymbol{\Gamma}_\mathbf{u}'\mathbf{Y}$. The matrix $\boldsymbol{\Gamma}_\mathbf{u}$ is used to form a basis of the space orthogonal to the direction $\mathbf{u}$. Then the $\boldsymbol{\tau}$th directional quantile of $\mathbf{Y}$ is a hyperplane $\lambda_{\boldsymbol{\tau}} = \{\mathbf{y} \in \Re^k : \mathbf{u}'\mathbf{y} = \beta_{\boldsymbol{\tau}}'\boldsymbol{\Gamma}_\mathbf{u}'\mathbf{y} + \alpha_{\boldsymbol{\tau}}\}$ where

$$(\alpha_{\boldsymbol{\tau}}, \beta_{\boldsymbol{\tau}}) \in \underset{a,\mathbf{b}}{argmin}\, E[\rho_\tau(\mathbf{Y}_\mathbf{u} - \mathbf{b}'\mathbf{Y}_\mathbf{u}^\perp - a)].$$

Denote $\mathbf{X} \in \Re^p$ to be random covariates. Define $\Psi(a, \mathbf{b}) = E[\rho_\tau(\mathbf{Y}_\mathbf{u} - \mathbf{b}_\mathbf{y}'\mathbf{Y}_\mathbf{u}^\perp - \mathbf{b}_\mathbf{x}'\mathbf{X} - a)]$.

The $\boldsymbol{\tau}$th quantile regression of $\mathbf{Y}$ on $\mathbf{X}$ (and an intercept) is $\lambda_{\boldsymbol{\tau}} = \{\mathbf{y} \in \Re^k : \mathbf{u}'\mathbf{y} = \beta'_{\boldsymbol{\tau}\mathbf{y}}\boldsymbol{\Gamma}'_{\mathbf{u}}\mathbf{y} + \beta'_{\boldsymbol{\tau}\mathbf{x}}\mathbf{X} + \alpha_{\boldsymbol{\tau}}\}$ where

$$(\alpha_{\boldsymbol{\tau}}, \beta_{\boldsymbol{\tau}}) = (\alpha_{\boldsymbol{\tau}}, \beta_{\boldsymbol{\tau}\mathbf{y}}, \beta_{\boldsymbol{\tau}\mathbf{x}}) \in \underset{a,\mathbf{b_y},\mathbf{b_x}}{argmin}\, \Psi(a, \mathbf{b}). \tag{3.7}$$

It is clear that the definition of the location case is embedded in definition (3.7) where $\mathbf{b_x}$ and $\mathbf{X}$ are of null dimension. Note that $\beta_{\boldsymbol{\tau}}$ is a function of $\boldsymbol{\Gamma}_{\mathbf{u}}$. This relationship is of little importance, the uniqueness of $\beta'_{\boldsymbol{\tau}}\boldsymbol{\Gamma}'_{\mathbf{u}}$ is of greater interest; which is unique under assumption 3.2 presented in the next section.

Any given quantile hyperplane, $\lambda_{\boldsymbol{\tau}}$, separates $\mathbf{Y}$ into two halfspaces, commonly referred to as regions. An open lower halfspace quantile halfspace,

$$H^-_{\boldsymbol{\tau}} = H^-_{\boldsymbol{\tau}}(\alpha_{\boldsymbol{\tau}}, \beta_{\boldsymbol{\tau}}) = \{y \in \Re^k : \mathbf{u}'\mathbf{y} < \beta'_{\boldsymbol{\tau}\mathbf{y}}\boldsymbol{\Gamma}'_{\mathbf{u}}\mathbf{y} + \beta'_{\boldsymbol{\tau}\mathbf{x}}\boldsymbol{\Gamma}'_{\mathbf{u}}\mathbf{X} + \alpha_{\boldsymbol{\tau}}\}, \tag{3.8}$$

and a closed upper quantile halfspace,

$$H^+_{\boldsymbol{\tau}} = H^+_{\boldsymbol{\tau}}(\alpha_{\boldsymbol{\tau}}, \beta_{\boldsymbol{\tau}}) = \{y \in \Re^k : \mathbf{u}'\mathbf{y} \geq \beta'_{\boldsymbol{\tau}\mathbf{y}}\boldsymbol{\Gamma}'_{\mathbf{u}}\mathbf{y} + \beta'_{\boldsymbol{\tau}\mathbf{x}}\boldsymbol{\Gamma}'_{\mathbf{u}}\mathbf{X} + \alpha_{\boldsymbol{\tau}}\}. \tag{3.9}$$

Under certain conditions, a distribution $\mathbf{Y}$ can be fully characterized by a family of hyperplanes $\Lambda = \{\lambda_{\boldsymbol{\tau}} : \boldsymbol{\tau} = \tau\mathbf{u} \in \mathcal{B}^k\}$ (Kong and Mizera, 2012, Theorem 5).[10] There are two subfamilies: a fixed-$\mathbf{u}$ subfamily, $\Lambda_{\mathbf{u}} = \{\lambda_{\boldsymbol{\tau}} : \boldsymbol{\tau} = \tau\mathbf{u}, \tau \in (0, 1)\}$, and a fixed-$\tau$ subfamily, $\Lambda_\tau = \{\lambda_{\boldsymbol{\tau}} : \boldsymbol{\tau} = \tau\mathbf{u}, \mathbf{u} \in \mathcal{S}^{k-1}\}$. The fixed-$\tau$ subfamily generates a fixed-τ region. The τ-quantile regression region is defined as

$$R(\tau) = \bigcap_{\mathbf{u}\in\mathcal{S}^{k-1}} \cap\{H^+_{\boldsymbol{\tau}}\}, \tag{3.10}$$

[10]The conditions required are the directional quantile envelopes of the probability distribution of $\mathbf{Y}$ with contiguous support have smooth boundaries for every $\tau \in (0, 0.5)$

where $\cap\{H_{\boldsymbol{\tau}}^{+}\}$ is the intersection over $H_{\boldsymbol{\tau}}^{+}$ if (3.7) is not unique. The boundary of $R(\tau)$ is called the τ quantile regression contour.

The boundary has a strong connection to Tukey (i.e. halfspace) depth contours. A depth function is a multivariate notion of centrality of an observation. Consider the set of all hyperplanes in $\Re^k$ that pass through some fixed point $\mathbf{y} \in \Re^k$. The Tukey depth of $\mathbf{y}$ is the minimum percentage of observations separated by all hyperplanes passing through $\mathbf{y}$. Hallin et al. (2010) show the fixed-τ region is equivalent to the Tukey (or halfspace) depth region.[11] This is advantageous because previous numerical approaches to Tukey depth contours were computationally expensive, however computation of directional quantiles is relatively easy.

If $\mathbf{Y}$ (or $\mathbf{Y}$ and $\mathbf{X}$ for the regression case) is absolutely continuous with respect to Lebesgue measure, has connected support and finite first moments then $(\alpha_\tau, \beta_\tau)$ and λ_τ are unique (Paindaveine and Šiman, 2011). This is assumption 2, which is stated formally in the next section.[12] Under this assumption the 'subgradient conditions' required for consistency are well defined. Further, $\Psi(a, \mathbf{b})$ is convex and continuously differentiable with respect to a and $\mathbf{b}$. The target parameters $(\alpha_{\boldsymbol{\tau}0}, \beta_{\boldsymbol{\tau}0})$ are defined as the parameters that satisfy the two subgradient conditions:

$$\left.\frac{\partial \Psi(a, \mathbf{b})}{\partial a}\right|_{\alpha_{\boldsymbol{\tau}0}, \beta_{\boldsymbol{\tau}0}} = Pr(\mathbf{Y_u} - \beta'_{\boldsymbol{\tau}\mathbf{y}0}\mathbf{Y}_{\mathbf{u}}^{\perp} - \beta'_{\boldsymbol{\tau}\mathbf{x}0}\mathbf{X} - \alpha_{\boldsymbol{\tau}0} \leq 0) - \tau = 0 \tag{3.11}$$

and

$$\left.\frac{\partial \Psi(a, \mathbf{b})}{\partial \mathbf{b}}\right|_{\alpha_{\boldsymbol{\tau}0}, \beta_{\boldsymbol{\tau}0}} = E[[\mathbf{Y}_{\mathbf{u}}^{\perp\prime}, \mathbf{X}']' 1_{(\mathbf{Y_u} - \beta'_{\boldsymbol{\tau}\mathbf{y}0}\mathbf{Y}_{\mathbf{u}}^{\perp} - \beta'_{\boldsymbol{\tau}\mathbf{x}0}\mathbf{X} - \alpha_{\boldsymbol{\tau}0} \leq 0)}] - \tau E[[\mathbf{Y}_{\mathbf{u}}^{\perp\prime}, \mathbf{X}']'] = \mathbf{0}_{k+p-1}. \tag{3.12}$$

The first condition can be equivalently written as $Pr(\mathbf{Y} \in H_{\boldsymbol{\tau}}^{-}) = \tau$ which maintains the

[11] Mathematically, the Tukey (or halfspace) depth of $\mathbf{y}$ with respect to probability distribution P is defined as $HD(\mathbf{y}, P) = \inf\{P[H] : H \text{ is a closed halfspace containing } \mathbf{y}\}$. Then the Tukey halfspace depth region is defined as $D(\tau) = \{\mathbf{y} \in \Re^k : HD(\mathbf{y}, P) \geq \tau\}$. Hallin et al. (2010) show $R(\tau) = D(\tau)$ for all $\tau \in [0, 1)$.

[12] This assumption can be weakened to only requiring moments to exist for $\mathbf{X}$ by using an alternative but equivalent definition of (3.7) based projection quantiles.

idea of a quantile partitioning the support into two sets, one with probability τ and one with probability $(1-\tau)$. The second condition can be written as

$$\tau = \frac{E[\mathbf{Y}^{\perp}_{\mathbf{u}i} 1_{(\mathbf{Y} \in H^{-}_{\boldsymbol{\tau}})}]}{E[\mathbf{Y}^{\perp}_{\mathbf{u}i}]} \text{ for all } i \in \{1, ..., k\}$$
$$\tau = \frac{E[\mathbf{X}_i 1_{(\mathbf{Y} \in H^{-}_{\boldsymbol{\tau}})}]}{E[\mathbf{X}_i]} \text{ for all } i \in \{1, ..., p\}$$

This condition can be interpreted as the probability mass center in the lower halfspace for the orthogonal response is $\tau \cdot 100\%$ that of the probability mass center in the entire space. Likewise, the probability mass center in the lower halfspace for the covariates is $\tau \cdot 100\%$ that of the probability mass center in the entire space.

Note $E[[\mathbf{Y}^{\perp\prime}_{\mathbf{u}}, \mathbf{X}']'] = E[[\mathbf{Y}^{\perp\prime}_{\mathbf{u}}, \mathbf{X}']' 1_{(\mathbf{Y} \in H^{+}_{\boldsymbol{\tau}})}] + E[[\mathbf{Y}^{\perp\prime}_{\mathbf{u}}, \mathbf{X}']' 1_{(\mathbf{Y} \in H^{-}_{\boldsymbol{\tau}})}]$, then the second condition can be written as

$$diag(\boldsymbol{\Gamma}'_{\mathbf{u}}, \mathbf{I}_p) \left[\frac{1}{1-\tau} E[[\mathbf{Y}', \mathbf{X}']' 1_{(\mathbf{Y} \in H^{+}_{\boldsymbol{\tau}})}] - \frac{1}{\tau} E[[\mathbf{Y}', \mathbf{X}']' 1_{(\mathbf{Y} \in H^{-}_{\boldsymbol{\tau}})}] \right] = \mathbf{0}_{k+p-1}.$$

The first $k-1$ components,

$$\boldsymbol{\Gamma}'_{\mathbf{u}} \left[\frac{1}{1-\tau} E[\mathbf{Y} 1_{(\mathbf{Y} \in H^{+}_{\boldsymbol{\tau}})}] - \frac{1}{\tau} E[\mathbf{Y} 1_{(\mathbf{Y} \in H^{-}_{\boldsymbol{\tau}})}] \right] = \mathbf{0}_{k-1},$$

show $\frac{1}{1-\tau} E[\mathbf{Y} 1_{(\mathbf{Y} \in H^{+}_{\boldsymbol{\tau}})}] - \frac{1}{\tau} E[\mathbf{Y} 1_{(\mathbf{Y} \in H^{-}_{\boldsymbol{\tau}})}]$ is orthogonal to $\boldsymbol{\Gamma}'_{\mathbf{u}}$ and thus, is parallel to $\mathbf{u}$. This states that the difference of the weighted probability mass centers of the two spaces is parallel to $\mathbf{u}$.

Figure 3.1 shows an example of these subgradient conditions with $1,000$ draws from $\mathbf{Y}$ when $\mathbf{Y}$ is distributed independently over the uniform unit square centered on $(0, 0)$. The directional vector is $\mathbf{u} = (1/\sqrt{2}, 1/\sqrt{2})$, which is the orange 45° degree arrow pointing to the top right. The depth is $\tau = 0.2$. The hyperplane $\lambda_{\boldsymbol{\tau}}$ is the red dotted line going from the top left to the bottom right. The lower quantile region $H^{-}_{\boldsymbol{\tau}}$ are the red dots lying below

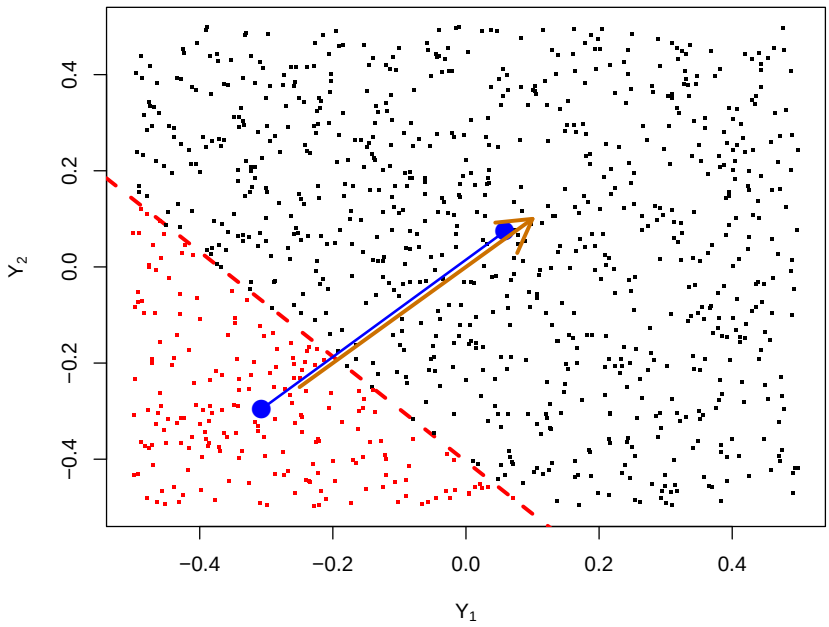

Figure 3.1: Lower quantile halfspace for $u = (1/\sqrt{2}, 1/\sqrt{2})$ and $\tau = 0.2$

$\lambda_{\boldsymbol{\tau}}$. The upper quantile region $H^{+}_{\boldsymbol{\tau}}$ are the black dots lying above $\lambda_{\boldsymbol{\tau}}$. The probability mass centers of the lower and upper quantile regions are represented by the solid blue dots in their respective regions. The first subgradient condition states that 20% of all points are red. The second subgradient condition states that the line joining the two probability mass centers is parallel to $\mathbf{u}$.

Figure 3.2 shows an example of fixed-τ regions (left) and fixed-$\mathbf{u}$ (right) halfspaces using the same simulated data as above. The left plot shows fixed-τ quantile upper halfspace intersections of 32 equally spaced directions on the unit circle for $\tau = 0.2$. The points on the boundary are all the Tukey depth points whose depth is 0.2. All the points within the shaded blue region have a Tukey depth greater than or equal to $\tau = 0.2$ and all points outside the shaded blue region have Tukey depth less than $\tau = 0.2$.

The right plot of figure 3.2 plot shows 13 quantile hyperplanes $\lambda_{\boldsymbol{\tau}}$ for a fixed $\mathbf{u} = (1/\sqrt{2}, 1/\sqrt{2})$

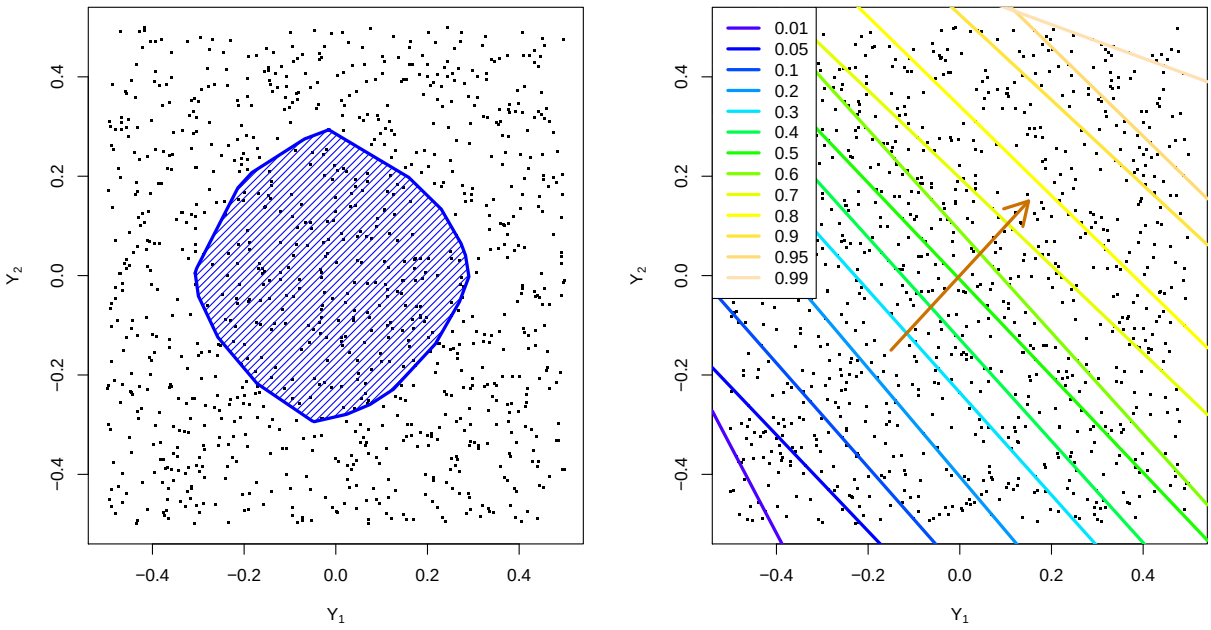

Figure 3.2: Example of a fixed-τ region and fixed-$\mathbf{u}$ halfspaces. Left, fixed $\tau = 0.2$ quantile region. Right, fixed $u = (1/\sqrt{2}, 1/\sqrt{2})$ quantile halfspaces.

for various τ (provided in the legend). The orange arrow shows the direction vector $\mathbf{u}$. The legend gives the value of τ used for each hyperplane. The hyperplanes split the square such that $\tau \cdot 100\%$ of all points lie below the hyperplanes. Note the hyperplanes do not need to be orthogonal to $\mathbf{u}$. However, the weighted probably mass centers (not shown) are parallel to $\mathbf{u}$.

Figure 3.3 shows an example of a fixed-τ regression tube through a random uniform pyramid. The left plot is a 3 dimensional scatter plot of the uniform pyramid.[13] The right plot shows the fixed-τ regression tube of Y_1 and Y_2 regressed on Y_3 with cross-section cuts at $Y_3 \in \{0, 0.15, 0.3\}$. As Y_3 increases the tube travels from the base to the tip of the pyramid. This causes the tube to pinch as the Y_3 increases. As in the one dimensional regression case, the regression tubes are susceptible to quantile crossing. Meaning if one were to trace out

[13] A uniform pyramid is a regular right pyramid where, for a fixed ϵ, every ϵ-ball contained within the pyramid has the same probability mass. The measure is normalized to one.

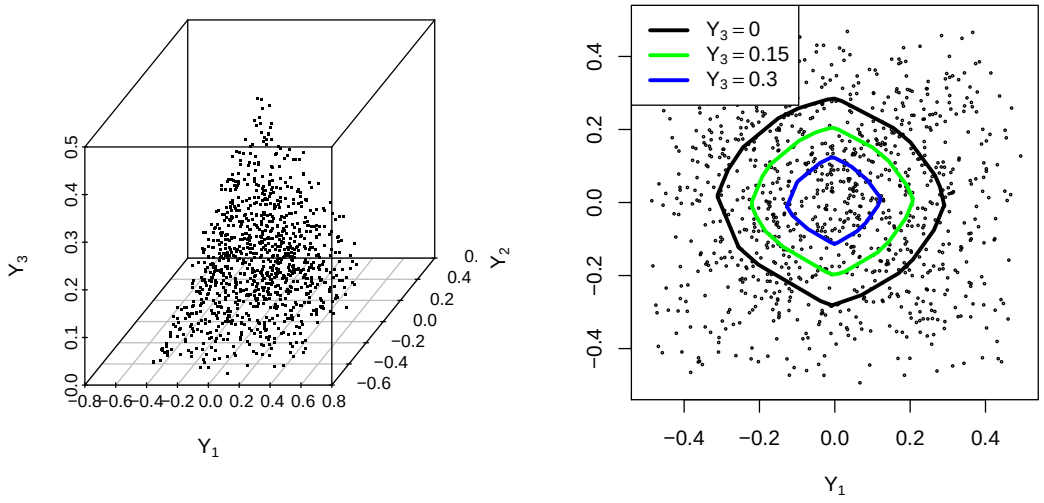

Figure 3.3: Example of a fixed-τ regression tube through a uniform pyramid. Left, a random uniform regular pyramid. Right, three slices of a fixed $\tau = 0.2$ regression tube.

the entire regression tube along Y_3 for a given τ and $\tau^\dagger > \tau$, the regression tube for $\tau^\dagger$ might not be contained in the one for τ for all Y_3.

3.2.1 Bayesian multiple-output quantile regression

The Bayesian approach assumes

$$\mathbf{Y_u}|\mathbf{Y_u^\perp}, \mathbf{X}, \alpha_{\boldsymbol{\tau}}, \beta_{\boldsymbol{\tau}} \sim ALD(\alpha_{\boldsymbol{\tau}} + \beta'_{\boldsymbol{\tau}\mathbf{y}}\mathbf{Y_u^\perp} + \beta'_{\boldsymbol{\tau}\mathbf{x}}\mathbf{X}, 1, \tau)$$

whose density is

$$f_{\boldsymbol{\tau}}(\mathbf{Y}|X, \alpha_{\boldsymbol{\tau}}, \beta_{\boldsymbol{\tau}}, \sigma_{\boldsymbol{\tau}}) = \frac{\tau(1-\tau)}{\sigma_{\boldsymbol{\tau}}} exp(-\frac{1}{\sigma_{\boldsymbol{\tau}}}\rho_\tau(\mathbf{Y} - \alpha_{\boldsymbol{\tau}} - \beta'_{\boldsymbol{\tau}\mathbf{y}}\mathbf{Y_u^\perp} - \beta'_{\boldsymbol{\tau}\mathbf{x}}\mathbf{X}))$$

and then chooses a prior $\Pi_{\boldsymbol{\tau}}(\alpha_{\boldsymbol{\tau}}, \beta_{\boldsymbol{\tau}})$ on the space $(\alpha_{\boldsymbol{\tau}}, \beta_{\boldsymbol{\tau}}) \in \Theta_{\boldsymbol{\tau}} \subset \Re^{k+p}$. The ALD distributional assumption likely does not represent the data generating process and is thus a

misspecified distribution. However, as more observations are obtained the posterior probability mass concentrates around neighborhoods of $(\alpha_{\boldsymbol{\tau}0}, \beta_{\boldsymbol{\tau}0})$, where $(\alpha_{\boldsymbol{\tau}0}, \beta_{\boldsymbol{\tau}0})$ satisfies (3.11) and (3.12). Theorem 3.1 shows this posterior consistency.

Denote the ith observation of the jth component of $\mathbf{Y}$ to be $\mathbf{Y}_{ij}$ and the ith observation of the lth covariate of $\mathbf{X}$ to be $\mathbf{X}_{il}$. The assumptions used are below.

Assumption 3.1. *The observations* $(\mathbf{Y}_i, \mathbf{X}_i)$ *are i.i.d. with true measure* $\mathbf{P}_0$ *for* $i \in \{1, 2, ..., n, ...\}$.

The density of $\mathbf{P}_0$ is denoted p_0. Assumption 3.1 states the observations are independent. This still allows for dependence among the components within a given observation.

The next assumption assures that the population parameters,$(\alpha_{\boldsymbol{\tau}0}, \beta_{\boldsymbol{\tau}0})$, are well defined by assuring the subgradient conditions exist and are unique.[14]

Assumption 3.2. *The measure of* $(\mathbf{Y}_i, \mathbf{X}_i)$ *is continuous with respect to Lebesgue measure, has connected support and admits finite first moments, for all* $i \in \{1, 2, ..., n, ...\}$.

The next assumption describes the prior.

Assumption 3.3. *The prior,* $\Pi_{\boldsymbol{\tau}}(\cdot)$*, has positive measure for every open neighborhood of* $(\alpha_{\boldsymbol{\tau}0}, \beta_{\boldsymbol{\tau}0})$ *and is*

a) proper, or

b) improper but admits a proper posterior.

Case b includes the Lebesgue measure on $\Re^{k+p}$ (i.e. flat prior) as a special case (Yu and Moyeed, 2001). Assumption 3.3 is satisfied using the prior suggested in section 3.3 for the Gibbs sampler.

[14] It is likely this assumption can be weakened (Serfling and Zuo, 2010)

The next assumption bounds the covariates and response variables.

Assumption 3.4. *There exists a* $c_x > 0$ *such that* $|\mathbf{X}_{i,l}| < c_x$ *for all* $l \in \{1, 2, ..., p\}$ *and all* $i \in \{1, 2, ..., n, ...\}$. *There exists a* $c_y > 0$ *such that* $|\mathbf{Y}_{i,j}| < c_y$ *for all* $j \in \{1, 2, ..., k\}$ *and all* $i \in \{1, 2, ..., n, ...\}$. *There exists a* $c_\Gamma > 0$ *such that* $\sup_{i,j} |[\mathbf{\Gamma_u}]_{i,j}| < c_\Gamma$.

The restriction on $\mathbf{X}$ is fairly mild in application, any given dataset will satisfy these restrictions. Further $\mathbf{X}$ can be controlled by the researcher in some situations (e.g. experimental environments). The restriction on $\mathbf{Y}$ is in conflict of the quantile regression attitude to remain agnostic about the distributional of the response. However, like $\mathbf{X}$, any given dataset will satisfy this restriction. The assumption on $\mathbf{\Gamma_u}$ is innocuous since $\mathbf{\Gamma_u}$ is chosen by the researcher, it is easy to choose such that all components are finite.

The next assumption ensures the Kullback Leibler minimizer is well defined.

Assumption 3.5. $E \log\left(\frac{p_0(\mathbf{Y}_i, \mathbf{X}_i)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|X_i, \alpha, \beta, 1)}\right) < \infty$ *for all* $i \in \{1, 2, ..., n, ...\}$.

The next assumption is to ensure the orthogonal response and covariate vectors are not degenerate.

Assumption 3.6. *There exist vectors* $\epsilon_Y > \mathbf{0}_{k-1}$ *and* $\epsilon_X > \mathbf{0}_p$ *such that*

$$Pr(\mathbf{Y}^{\perp}_{\mathbf{u}ij} > \epsilon_{Yj}, \mathbf{X}_{il} > \epsilon_{Xl}, \forall j \in \{1, ..., k-1\}, \forall l \in \{1, ..., p\}) = c_p \notin \{0, 1\}.$$

This assumption can always be satisfied with a simple location shift as long as each variable takes on two different values with positive joint probability. Let $U \subseteq \Theta$, define the posterior probability of U to be

$$\Pi_{\boldsymbol{\tau}}(U|(\mathbf{Y}_1, \mathbf{X}_1), (\mathbf{Y}_2, \mathbf{X}_2), ..., (\mathbf{Y}_n, \mathbf{X}_n)) = \frac{\int_U \prod_{i=1}^n \frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\mathbf{X}_i, \alpha_{\boldsymbol{\tau}}, \beta_{\boldsymbol{\tau}}, \sigma_{\boldsymbol{\tau}})}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\mathbf{X}_i, \alpha_{\boldsymbol{\tau}0}, \beta_{\boldsymbol{\tau}0}, \sigma_{\boldsymbol{\tau}0})} d\Pi_{\boldsymbol{\tau}}(\alpha_{\boldsymbol{\tau}}, \beta_{\boldsymbol{\tau}})}{\int_\Theta \prod_{i=1}^n \frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\mathbf{X}_i \alpha_{\boldsymbol{\tau}}, \beta_{\boldsymbol{\tau}}, \sigma_{\boldsymbol{\tau}})}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\mathbf{X}_i, \alpha_{\boldsymbol{\tau}0}, \beta_{\boldsymbol{\tau}0}, \sigma_{\boldsymbol{\tau}0})} d\Pi_{\boldsymbol{\tau}}(\alpha_{\boldsymbol{\tau}}, \beta_{\boldsymbol{\tau}})}.$$

The main theorem of the paper can now be stated.

Theorem 3.1. *Suppose assumptions 3.1, 3.2, 3.3a, 3.4 and 3.6 hold or assumptions 3.1, 3.2, 3.3b, 3.4, 3.5 and 3.6. Let* $U = \{(\alpha_{\boldsymbol{\tau}}, \beta_{\boldsymbol{\tau}}) : |\alpha_{\boldsymbol{\tau}} - \alpha_{\boldsymbol{\tau}0}| < \Delta, |\beta_{\boldsymbol{\tau}} - \beta_{\boldsymbol{\tau}0}| < \Delta \mathbf{1}_{k-1}\}$. *Then* $\lim_{n\to\infty} \Pi_{\boldsymbol{\tau}}(U^c|(\mathbf{Y}_1, \mathbf{X}_1), ..., (\mathbf{Y}_n, \mathbf{X}_n)) = 0$ *a.s.* $[\mathbf{P}_0]$.

The proof is presented in the appendix. The strategy of the proof follows very closely to the strategy used in the conditional one dimension case (Sriram et al., 2013). First construct an open set U_n containing $(\alpha_{\boldsymbol{\tau}0}, \beta_{\boldsymbol{\tau}0})$ for all n that converges to $(\alpha_{\boldsymbol{\tau}0}, \beta_{\boldsymbol{\tau}0})$, the target parameters. Define $B_n = \Pi_{\boldsymbol{\tau}}(U_n^c|(\mathbf{Y}_1, \mathbf{X}_1), ..., (\mathbf{Y}_n, \mathbf{X}_n))$. To show convergence of B_n to $B = 0$ almost surely, it is sufficient to show $\lim_{n\to\infty} \sum_{i=1}^n E[|B_n - B|^d] < \infty$ for some $d > 0$, using the Markov inequality and Borel-Cantelli lemma. The Markov inequality states if $B_n - B \geq 0$ then for any $d > 0$

$$Pr(|B_n - B| > \epsilon) \leq \frac{E[|B_n - B|^d]}{\epsilon^d}$$

for any $\epsilon > 0$. The Borel-Cantelli lemma states

$$\text{if } \lim_{n\to\infty} \sum_{i=1}^n Pr(|B_n - B| > \epsilon) < \infty \text{ then } Pr(\limsup_{n\to\infty} |B_n - B| > \epsilon) = 0.$$

Thus by Markov inequality

$$\sum_{i=1}^n Pr(|B_n - B| > \epsilon) \leq \sum_{i=1}^n \frac{E[|B_n - B|^d]}{\epsilon^d}.$$

Since $\lim_{n\to\infty} \sum_{i=1}^n E[|B_n - B|^d] < \infty$ then $\lim_{n\to\infty} \sum_{i=1}^n Pr(|B_n - B| > \epsilon) < \infty$. By Borel-Cantelli

$$Pr(\limsup_{n\to\infty} |B_n - B| > \epsilon) = 0.$$

To show $\lim_{n\to\infty} \sum_{i=1}^n E[|B_n - B|^d] < \infty$, I create a set G_n where $(\alpha_{\tau 0}, \beta_{\tau 0}) \notin G_n$. Within this set I show the expectation of the numerator is less than $e^{-2n\delta}$ and the expectation of the

denominator is greater than $e^{-n\delta}$ for some $\delta > 0$. Then the expected value of the posterior is less than $e^{-n\delta}$, which is summable. This exposition is a simplification from what is shown in the formal proof.

3.2.2 Choice of prior

A new model is estimated for each unique $\boldsymbol{\tau}$ and thus a prior is needed for each one. This might seem like there is an overwhelming amount of ex-ante elicitation required. However, simplifications can be made to make elicitation easier. If the prior is centered over H_0 : $\beta_{\boldsymbol{\tau}} = \mathbf{0}_{k+p-1}$ for all $\boldsymbol{\tau}$ then the implied ex-ante belief is $\mathbf{Y}$ has spherical Tukey contours and $\mathbf{X}$ has no relation with $\mathbf{Y}$.[15] Under this hypothesis ($H_0 : \beta_{\boldsymbol{\tau}} = \mathbf{0}_{k+p-1}$ for all $\boldsymbol{\tau}$), $\alpha_{\boldsymbol{\tau}}$ is the shortest euclidean distance of the τth Tukey contour from the Tukey median. Since the contours are spherical, the distance is the same for all $\mathbf{u}$. The variance of the prior expresses the researcher's confidence in the hypothesis of spherical Tukey contours. A large prior variance allows for large departures from H_0. If one is willing to accept joint normality of $\theta_{\boldsymbol{\tau}} = (\alpha_{\boldsymbol{\tau}}, \beta_{\boldsymbol{\tau}})$ then a Gibbs sampler can be used. The sampler is presented in the next section.

Further, if data is being collected and analyzed in real time, then the prior of the current analysis can be centered over the estimates from the previous analysis and the variance of the prior is the willingness the researcher is to allow for departures from the previous analysis.

Arbitrary priors not centered over 0 require a more detailed discussion. I will restrict to

[15] A sufficient condition for a density to have spherical Tukey contours is for the PDF to have spherical density contours and that its PDF (with a multivariate argument, $\mathbf{Y}$) can be written as a monotonically decreasing function of the inner product of the multivariate argument (i.e. $\mathbf{Y}'\mathbf{Y}$) (Dutta et al., 2011). This condition is satisfied for the location family for the standard multivariate Normal, T and Cauchy. The distance of the Tukey median and the τth Tukey contour for the multivariate standard normal is $\Phi^{-1}(1-\tau)$. Another distribution with spherical Tukey contours is the uniform hyperball. The distance of the Tukey median and the τth Tukey contour for the uniform hyperball is the r such that $arcsin(r) + r\sqrt{1-r^2} = \pi(0.5-\tau)$. This function is invertible for $r \in (0,1)$ and $\tau \in (0,.5)$ and can be computed using numerical approximations (Rousseeuw and Ruts, 1999).

the 2 dimensional case ($k = 2$). There are two ways to think of appropriate priors for $\theta_{\boldsymbol{\tau}} = (\alpha_{\boldsymbol{\tau}}, \beta_{\boldsymbol{\tau}})$. The first is a direct approach to think of, $\theta_{\boldsymbol{\tau}}$ as the slope of $\mathbf{Y_u}$ against $\mathbf{Y_u^\perp}$, $\mathbf{X}$ and an intercept. The second approach is to think of it in terms of the implied prior of $\phi_{\boldsymbol{\tau}} = \phi_{\boldsymbol{\tau}}(\theta_{\boldsymbol{\tau}})$ as the slope of Y_2 against Y_1, $\mathbf{X}$ and an intercept. The second approach is presented in the appendix.

In the direct approach the parameters relate directly to the subgradient conditions (3.11) and (3.12).[16] Under the hypothesis $H_0 : \beta_{\tau\mathbf{y}} = 0$ the hyperplane $\lambda_{\boldsymbol{\tau}}$ is orthogonal to $\mathbf{u}$ (and thus $\lambda_{\boldsymbol{\tau}}$ is parallel to $\Gamma_{\mathbf{u}}$). As $|\beta_{\boldsymbol{\tau}\mathbf{y}}| \to \infty$, $\lambda_{\boldsymbol{\tau}}$ converges to $\mathbf{u}$ monotonically.[17] A δ unit increase in $\beta_{\boldsymbol{\tau}y}$ tilts the $\lambda_{\boldsymbol{\tau}}$ hyperplane.[18] The direction of the tilt is determined by the vectors $\mathbf{u}$ and $\Gamma_{\mathbf{u}}$ and the sign of δ. The vectors $\mathbf{u}$ and $\Gamma_{\mathbf{u}}$ always form 2 angles: a 90° and 270° angle. For positive δ, the hyperplane travels monotonically through the triangle formed by the 90°. For negative δ the hyperplane travels monotonically in the opposite direction.

The value of $|\alpha_{\boldsymbol{\tau}}|$ is the euclidean distance from the Tukey median to the point where $\lambda_{\boldsymbol{\tau}}$ intersects $\mathbf{u}$. A δ unit increase in $\alpha_{\boldsymbol{\tau}}$ results in a parallel shift in the hyperplane $\lambda_{\boldsymbol{\tau}}$ by $\frac{\delta}{u_2 - \beta_{\boldsymbol{\tau}\mathbf{y}} u_2^\perp}$ units.

[16] The vector $\mathbf{Y_u}$ is the scalar projection of $\mathbf{Y}$ in direction $\mathbf{u}$ and $\mathbf{Y_u^\perp}$ is the scalar projection of $\mathbf{Y}$ in the direction of the other (orthogonal) basis vectors.

[17] Monotonic meaning the angular distance between $\lambda_{\boldsymbol{\tau}}$ and $\mathbf{u}$ is always decreasing for strictly increasing or decreasing $\beta_{\boldsymbol{\tau}\mathbf{y}}$.

[18] Define $slope(\delta)$ to be the slope of the hyperplane when β is increased by δ. The slope of the new hyperplane is $slope(\delta) = (u_2 - (\beta + \delta)u_2^\perp)^{-1}(\delta u_1^\perp + (u_2 - \beta u_2^\perp)slope(0)$.

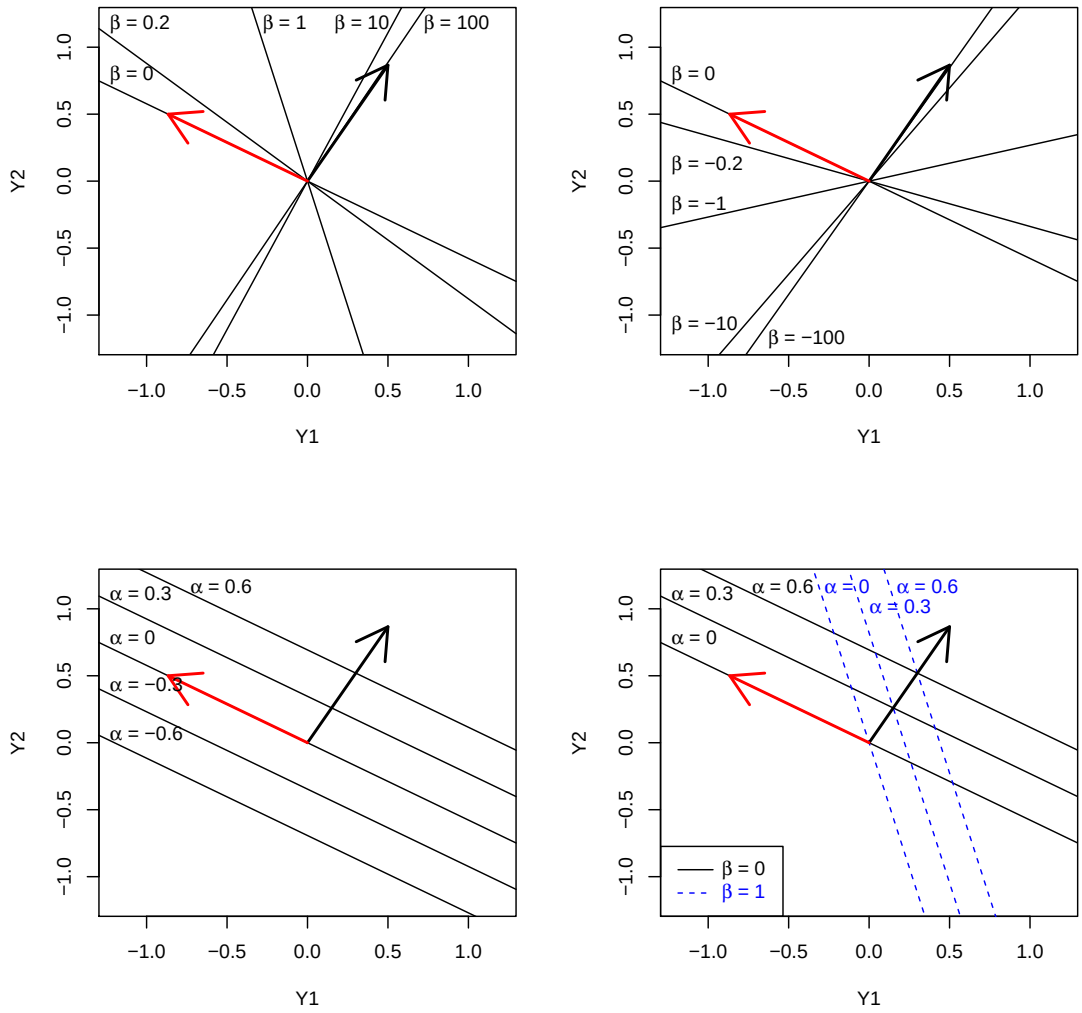

Figure 3.4: Hyperplanes from various hyperparameters ($\boldsymbol{\tau}$ subscript omitted). Top left, positive β. Top right, negative β. Bottom left, different αs. Bottom right, different αs and βs.

Figure 3.4 shows the prior hyperplanes from various hyperparameters. For all four plots $k = 2$, the directional vector is $\mathbf{u} = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$ and $\Gamma_{\mathbf{u}} = (-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$. The top left plot shows the hyperplanes for $\beta_{\boldsymbol{\tau}}$ increasing from 0 to 100 for fixed $\alpha_{\boldsymbol{\tau}} = 0$. At $\beta = 0$ the hyperplane is perpendicular to $\mathbf{u}$ as it increases the hyperplane travels counterclockwise until it becomes parallel to $\mathbf{u}$. The top right plot shows the hyperplanes for $\beta_{\boldsymbol{\tau}}$ decreasing from 0 to -100 for fixed $\alpha = 0$. At $\beta_{\boldsymbol{\tau}} = 0$ the hyperplane is perpendicular to $\mathbf{u}$ as it decreases the hyperplane

travels clockwise until it becomes parallel to $\mathbf{u}$. The bottom left plot shows the hyperplane for $\alpha_{\boldsymbol{\tau}}$ ranging from -0.6 to 0.6. The Tukey median can be thought of the point $(0, 0)$, then $|\alpha_{\boldsymbol{\tau}}|$ is the distance of the intersection of $\mathbf{u}$ and $\lambda_{\boldsymbol{\tau}}$ from the Tukey median.[19] For positive $\alpha_{\boldsymbol{\tau}}$ the hyperplanes are moving in the direction $\mathbf{u}$ and for negative $\alpha_{\boldsymbol{\tau}}$ the hyperplanes are moving in the direction $-\mathbf{u}$. The bottom right plot shows the hyperplanes for various $\alpha_{\boldsymbol{\tau}}$ and $\beta_{\boldsymbol{\tau}}$. The solid black hyperplanes are for $\beta_{\boldsymbol{\tau}} = 0$ and the dashed blue hyperplanes are for $\beta_{\boldsymbol{\tau}} = 1$ and $\alpha_{\boldsymbol{\tau}}$ takes on values 0, 0.3 and 0.6 for both values of $\beta_{\boldsymbol{\tau}}$. This plot confirms changes in $\alpha_{\boldsymbol{\tau}}$ result in parallel shifts of $\lambda_{\boldsymbol{\tau}}$ while $\beta_{\boldsymbol{\tau}}$ tilts $\lambda_{\boldsymbol{\tau}}$.

3.3 Computation

If one is willing to accept joint normality of the prior distribution for the parameters then estimation can be performed using the Gibbs sampler developed in Kozumi and Kobayashi (2011). The approach is to assume $\mathbf{Y}_{\mathbf{u}i} = \beta'_{\boldsymbol{\tau}\mathbf{y}}\mathbf{Y}^{\perp}_{\mathbf{u}i} + \beta'_{\boldsymbol{\tau}\mathbf{x}}\mathbf{X}_i + \alpha_{\boldsymbol{\tau}} + \epsilon_i$ where $\epsilon_i \overset{iid}{\sim} ALD(0, 1)$. The random component, ϵ_i, can be written as a mixture of a normal and exponential, $\epsilon_i = \eta W_i + \gamma\sqrt{W_i}U_i$ where $\eta = \frac{1-2\tau}{\tau(1-\tau)}$, $\gamma = \sqrt{\frac{2}{\tau(1-\tau)}}$, $W_i \overset{iid}{\sim} exp(1)$ and $U_i \overset{iid}{\sim} N(0, 1)$ are mutually independent (Kotz et al., 2001). Then $\mathbf{Y}_{\mathbf{u}i}|\mathbf{Y}^{\perp}_{\mathbf{u}i}, \mathbf{X}_i, W_i, \beta_{\boldsymbol{\tau}}, \alpha_{\boldsymbol{\tau}}$ is normally distributed. If the prior is $\theta_{\boldsymbol{\tau}} = (\alpha_{\boldsymbol{\tau}}, \beta_{\boldsymbol{\tau}}) \sim N(\mu_{\theta_{\boldsymbol{\tau}}}, \Sigma_{\theta_{\boldsymbol{\tau}}})$ then a Gibbs sampler can be used. The $m + 1$th MCMC draw is given by the following algorithm

1. Draw $W_i^{(m+1)} \sim W|\mathbf{Y}_{\mathbf{u}i}, \mathbf{Y}^{\perp}_{\mathbf{u}i}, \mathbf{X}_i, Z_i, \theta^{(m)}_{\boldsymbol{\tau}} \sim GIG(\frac{1}{2}, \hat{\delta}_i, \hat{\phi}_i)$ for $i \in \{1, ..., n\}$

2. Draw $\theta^{(m+1)}_{\boldsymbol{\tau}} \sim \theta_{\boldsymbol{\tau}}|\vec{\mathbf{Y}}_{\mathbf{u}}, \vec{\mathbf{Y}}^{\perp}_{\mathbf{u}}, \vec{\mathbf{X}}, \vec{Z}, \vec{W}^{(m+1)} \sim N(\hat{\theta}_{\boldsymbol{\tau}}, \hat{B}_{\boldsymbol{\tau}})$.

[19] The Tukey median does not exist in these plots since there is no data. If there was data, the point where $\mathbf{u}$ and $\Gamma_{\mathbf{u}}$ intersect would be the Tukey median.

where

$$\hat{\delta}_i = \frac{1}{\gamma^2}(\mathbf{Y}_{\mathbf{u}i} - \beta'^{(m)}_{\boldsymbol{\tau}\mathbf{y}}\mathbf{Y}^{\perp}_{\mathbf{u}i} - \beta'^{(m)}_{\boldsymbol{\tau}\mathbf{x}}\mathbf{X}_i - \alpha^{(m+1)}_{\boldsymbol{\tau}})^2$$

$$\hat{\phi}_i = 2 + \frac{\eta^2}{\gamma^2}$$

$$\hat{B}^{-1}_{\boldsymbol{\tau}} = B^{-1}_{\boldsymbol{\tau}0} + \sum_{i=1}^{n} \frac{[\mathbf{Y}^{\perp}_{\mathbf{u}i}, \mathbf{X}_i][\mathbf{Y}^{\perp}_{\mathbf{u}i}, \mathbf{X}_i]'}{\gamma^2 W_i}$$

$$\hat{\beta}_{\boldsymbol{\tau}} = \hat{B}_{\boldsymbol{\tau}}\left(B^{-1}_{\boldsymbol{\tau}0}\beta_{\boldsymbol{\tau}0} + \sum_{i=1}^{n} \frac{[\mathbf{Y}^{\perp}_{\mathbf{u}i}, \mathbf{X}_i]'(\mathbf{Y}_{\mathbf{u}i} - \eta W_i^{(m+1)})}{\gamma^2 W_i^{(m+1)}}\right)$$

and $GIG(\nu, a, b)$ is the Generalized Inverse Gamma distribution whose density is

$$f(x|\nu, a, b) = \frac{(b/a)^{\nu}}{2K_{\nu}(ab)} x^{\nu-1} exp(-\frac{1}{2}(a^2x^{-1} + b^2x)), x > 0, -\infty < \nu < \infty, a, b \geq 0$$

and $K_{\nu}(\cdot)$ is the modified Bessel function of the third kind. An efficient sampler of the Generalized Inverse Gamma distribution was developed in Dagpunar (1989). An implementation of the Gibbs sampler for R is provided in the package 'bayesQR' (Benoit et al., 2014). The sampler is geometrically ergodic and thus the MCMC standard error is finite and the MCMC central limit theorem is well defined (Khare and Hobert, 2012). This guarantees that after a long enough burn-in draws from this sampler are equivalent to random draws from the posterior.

3.4 Simulation

In this section I show the consistency of the procedure as well as show its robustness to violations of Assumption 3.4. Consistency is verified by checking for convergence of the subgradient conditions. I consider four DGPs

1. $\mathbf{Y} \sim Uniform\ Square$

2. $\mathbf{Y} \sim Uniform\ Triangle$

3. $\mathbf{Y} \sim N(\mu, \Sigma)$, where $\mu = \mathbf{0}_2$ and $\Sigma = \begin{bmatrix} 1 & 1.5 \\ 1.5 & 9 \end{bmatrix}$

4. $\mathbf{Y} = \mathbf{Z} + \begin{bmatrix} 0 \\ X \end{bmatrix}$ where $\begin{bmatrix} X \\ \mathbf{Z} \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_X \\ \mu_{\mathbf{Z}} \end{bmatrix}, \begin{bmatrix} \Sigma_{XX} & \Sigma_{X\mathbf{Z}} \\ \Sigma'_{X\mathbf{Z}} & \Sigma_{\mathbf{ZZ}} \end{bmatrix} \right)$,

$\Sigma_{XX} = 4$, $\Sigma_{X\mathbf{Z}} = \begin{bmatrix} 0 & 2 \end{bmatrix}$, $\Sigma_{\mathbf{ZZ}} = \begin{bmatrix} 1 & 1.5 \\ 1.5 & 9 \end{bmatrix}$, $\mu_X = 0$ and $\mu_{\mathbf{Z}} = \mathbf{0}_2$

The first DGP has corners at $(0,0), (0,1), (1,1), (1,0)$. The second DGP has corners at $(-1,0), (1,0), (0,\sqrt{3})$. DGPs 1,2 and 3 are location models and 4 is a regression model. DGPs 1 and 2 conform to all the assumptions on the data generating process. DGPs 3 and 4 are cases when Assumption 3.4 is violated. In DGP 4, the unconditional distribution of $\mathbf{Y}$ is $\mathbf{Y} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 1.5 \\ 1.5 & 17 \end{bmatrix} \right)$. To verify consistency I check for convergence of the subgradient conditions (3.11) and (3.12). Define $\hat{H}_{\boldsymbol{\tau}}$ to be the empirical lower halfspace where the parameters in (3.8) are replaced with their estimators. To check the first subgradient condition (3.11), I verify

$$\frac{1}{n} \sum_{i=1}^{n} 1_{(\mathbf{Y}_i \in \hat{H}_{\boldsymbol{\tau}})} \to \tau. \tag{3.13}$$

Since $\mathbf{Y}_{\mathbf{u}}$ is one dimension, computation of $1_{(\mathbf{Y}_i \in \hat{H}_{\boldsymbol{\tau}})}$ is simple. To check the second subgradient condition (3.12), I verify

$$\frac{1}{n} \sum_{i=1}^{n} \mathbf{Y}^{\perp}_{\mathbf{u}i} 1_{(\mathbf{Y}_i \in \hat{H}_{\boldsymbol{\tau}})} \to \tau E[\mathbf{Y}^{\perp}_{\mathbf{u}}] \tag{3.14}$$

and

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i 1_{(\mathbf{Y}_i\in\hat{H}_{\boldsymbol{\tau}})} \to \tau E[\mathbf{X}]. \tag{3.15}$$

Similar to the first subgradient condition, computation of $\mathbf{Y}^{\perp}_{\mathbf{u}i}1_{(\mathbf{Y}_i\in\hat{H}_{\boldsymbol{\tau}})}$ and $\mathbf{X}_i 1_{(\mathbf{Y}_i\in\hat{H}_{\boldsymbol{\tau}})}$ is simple. For DGPs 1-4, $E[\mathbf{Y}^{\perp}_{\mathbf{u}}] = \mathbf{0}_2$ and for DGP 4, $E[\mathbf{X}] = 0$.

Two directions are considered $\mathbf{u}$: $(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$ and $(0, 1)$. The first vector is a 45^o line between Y_2 and Y_1 in the positive quadrant. The second vector points vertically in the Y_2 direction. The sample sample sizes are $n = 100,\ 1,000$ and $10,000$. The depths are $\tau = 0.2$ and $\tau = 0.4$. The prior is $\theta_{\boldsymbol{\tau}} \sim N(\mu_{\theta_{\boldsymbol{\tau}}}, \Sigma_{\theta_{\boldsymbol{\tau}}})$ where $\mu_{\theta_{\boldsymbol{\tau}}} = \mathbf{0}_{k+p-1}$ and $\Sigma_{\theta_{\boldsymbol{\tau}}} = 1000\mathbf{I}_{k+p-1}$.

		Data Generating Process			
	n	1	2	3	4
Sub Grad 1	100	4.47e-02	2.91e-02	1.52e-02	1.75e-02
	1,000	5.44e-03	4.59e-03	2.48e-03	2.60e-03
	10,000	9.29e-04	8.66e-04	5.42e-04	5.12e-04
Sub Grad 2	100	6.34e-03	1.43e-02	4.34e-02	7.06e-02
	1,000	2.01e-03	3.29e-03	1.32e-02	2.05e-02
	10,000	5.82e-04	8.00e-04	3.59e-03	4.91e-03

Table 3.1: RMSE of subgradient conditions for $\mathbf{u} = (1/\sqrt{2}, 1/\sqrt{2})$

		Data Generating Process			
	n	1	2	3	4
Sub Grad 1	100	2.02e-02	1.89e-02	1.16e-02	1.36e-02
	1,000	3.38e-03	3.61e-03	1.96e-03	1.98e-03
	10,000	7.71e-04	9.32e-04	3.87e-04	4.68e-04
Sub Grad 2	100	9.74e-03	1.35e-02	2.59e-02	2.29e-02
	1,000	2.08e-03	3.24e-03	7.11e-03	6.51e-03
	10,000	6.15e-04	9.89e-04	2.01e-03	1.83e-03

Table 3.2: RMSE of subgradient conditions for $\mathbf{u} = (0, 1)$

Tables 3.1, 3.2 and 3.3 show the results from the simulation. Tables 3.1 and 3.2 show the Root Mean Square Error (RMSE) of (3.13) and (3.14). The first three rows show the RMSE for the first subgradient condition (3.13). The last three rows show the RMSE for the second subgradient condition (3.12). The second column, n, is the sample size. The

next five columns are the DGPs previously described. Table 3.1 is using directional vector $\mathbf{u} = (1/\sqrt{2}, 1/\sqrt{2})$ and Table 3.2 is using directional vector $\mathbf{u} = (0, 1)$. It is clear that as sample size increases the RMSEs are decreasing, showing the convergence of the subgradient conditions.

	Direction $\mathbf{u}$	
n	$(1/\sqrt{2}, 1/\sqrt{2})$	$(0, 1)$
100	5.17e-02	5.17e-02
1,000	1.41e-02	1.41e-02
10,000	3.90e-03	3.90e-03

Table 3.3: RMSE of regressor subgradient condition for DGP 4

Table 3.3 shows RMSE of the covariate for DGP 4 (3.15) for convergence of subgradient condition (3.12). The three rows show sample size and the two columns show direction. It is clear that as sample size increases the RMSEs are decreasing, showing convergence of the subgradient conditions.

3.5 Application

I apply the model to educational data collected from the Project STAR public access database. Project STAR was an experiment conducted on 11,600 students in 300 classrooms from 1985-1989 with interest of determining if reduced classroom size improved academic performance. Students and teachers were randomly selected in kindergarten to be in small (13-17 students) or large (22-26 students) classrooms.[20] The students then stayed in their assigned classroom size throughout the fourth grade. The outcome of the treatment was measured using reading and mathematics test scores that were given each year. This dataset has been analyzed many times before, see Finn and Achilles (1990); Folger and Breda (1989); Krueger (1999); Mosteller (1995); Word et al. (1990).[21] The studies performed analyses on

[20]Some large classrooms also had a teaching assistant, I do not consider those classrooms in this paper.

[21] Folger and Breda (1989) and Finn and Achilles (1990) were the first two published studies. Word et al. (1990) was the official report from the Tennessee State Department of Education. Mosteller (1995) provided

either univariate test score measures or on an average of math, reading and word recognition scores. Univariate analysis ignores important information about the relationship the mathematics and reading test scores might have with each other. Analysis on the the average of scores better accommodates joint effects but obscures the source of an effect. Using multiple-output quantile regression I can obtain inferences on the joint relationship between scores for the entire multivariate distribution (or several specified quantile subpopulations). My results agree with and strengthen all previous studies.

A student's outcome was measured using a standardized test called the Stanford Achievement Test (SAT) for mathematics and reading.[22] This paper compared the outcomes of small and large classrooms on the subset of first grade students resulting in a sample size of $n = 6,379$ (after removal of missing data). The results for other grades were similar.[23]

Define the vector $\mathbf{u} = (u_1, u_2)$, where u_1 is the math score dimension and u_2 is the reading score dimension. The $\mathbf{u}$ directions have an interpretation of relating how much relative importance the researcher wants to give to math or reading. Define $\mathbf{u}^{\perp} = (u_1^{\perp}, u_2^{\perp})$, where $\mathbf{u}^{\perp}$ is orthogonal to $\mathbf{u}$. The components $(u_1^{\perp}, u_2^{\perp})$ have no meaningful interpretation. Define $math_i$ to be the math score of student i and $reading_i$ to be the reading score of student i.

a review of the study and Krueger (1999) performed a rigorous econometric analysis focusing on validity.

[22]The test scores have a finite discrete support ranging from . Computationally, this does not effect the Bayesian estimates, however prevents asymptotically unique estimators. So I perturb each of the scores with a uniform(0,1) random variable. I would like to thank Brian Bucks for this idea.

[23]The data analysis in this paper is used to explain the concepts of Bayesian multiple-output quantile regression, not to provide rigorous causal econometric inferences In the later case, a thorough discussion of missing data would be necessary. For the same reason first grade scores were chosen. The first grade subset was best suited for pedagogy. This experiment has been analyzed by many other researchers.

The model is

$$\begin{aligned}
\mathbf{Y}_{\mathbf{u}i} &= math_i u_1 + reading_i u_2 \\
\mathbf{Y}^{\perp}_{\mathbf{u}i} &= math_i u_1^{\perp} + reading_i u_2^{\perp} \\
\mathbf{Y}_{\mathbf{u}i} &= \alpha_{\boldsymbol{\tau}} + \beta_{\boldsymbol{\tau}} \mathbf{Y}^{\perp}_{\mathbf{u}i} + \epsilon_i \\
\epsilon_i &\overset{iid}{\sim} ALD(0, 1, \tau) \\
\theta_{\boldsymbol{\tau}} = (\alpha_{\boldsymbol{\tau}}, \beta_{\boldsymbol{\tau}}) &\sim N(\mu_{\theta_{\boldsymbol{\tau}}}, \Sigma_{\theta_{\boldsymbol{\tau}}}).
\end{aligned}$$

Unless otherwise noted, $\mu_{\theta_{\boldsymbol{\tau}}} = \mathbf{0}_2$ and $\Sigma_{\theta_{\boldsymbol{\tau}}} = 1000\mathbf{I}_2$, meaning ex-ante knowledge is a weak belief that the joint distribution of math and reading has spherical Tukey contours.

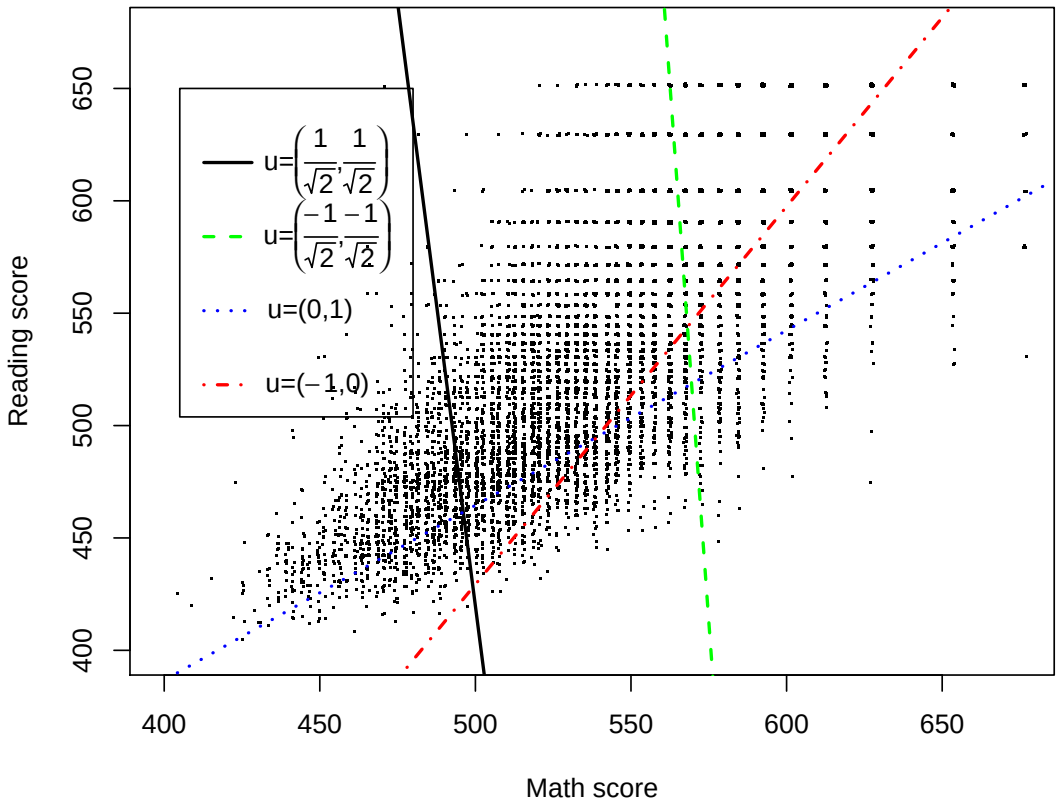

Figure 3.5: Various directional vectors for $\tau = 0.2$

The directional vectors, $\mathbf{u}$, are interpretable in the context of of this example. Figure 3.5 shows several hyperplanes for four different $\mathbf{u}$ directions with a fixed $\tau = 0.2$. The lower contour halfspace for $\mathbf{u} = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$ pointing 45° to the top right is interested in the halfspace with the $\tau \cdot 100\%$ students who performed the worst on the tests giving equal weight to math and reading. This $\mathbf{u}$ direction results in the solid black line and the lower quantile halfspace are all values that lie below it. Conversely, lower contour halfspace for $\mathbf{u} = (-\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}})$ pointing 225° to the bottom left is interested in the halfspace with the $\tau \cdot 100\%$ of students who performed the best on the tests giving equal weight to math and english. This $\mathbf{u}$ direction results in the dashed green line and the lower quantile halfspace are all values that lie to the right of it.

The lower contour halfspace for $\mathbf{u} = (0, 1)$ pointing 90° straight up is only interested in the worst performing $\tau \cdot 100\%$ of students for math. This $\mathbf{u}$ direction results in the dotted blue line and the lower quantile halfspace are all values that lie below it. The lower contour halfspace for $\mathbf{u} = (-1, 0)$ pointing 180° to the left is only interested in the best performing $\tau \cdot 100\%$ of students for reading. This $\mathbf{u}$ direction results in the dash-dot red line and the lower quantile halfspace are all values that lie below it.

Even though $\mathbf{u}$ can be interpreted as a weight vector, the slope of the hyperplanes are governed by the relative probability masses of the data. Note that the first two directions are 180° degrees of each other and their hyper planes are roughly parallel. This is not a requirement of the model. If it were, one might suspect that two orthogonal directions would result in orthogonal hyperplanes, but this is not that case. The second two directions are orthogonal but their hyperplanes are not. How the tilt is determined can better be understood with fixed-$\mathbf{u}$ hyperplanes, presented next.

Figure 3.6 are fixed-$\mathbf{u}$ contours for various τ along a fixed $\mathbf{u}$ direction. Two directions are used: $\mathbf{u} = (1/\sqrt{2}, 1/\sqrt{2})$ (left) and $\mathbf{u} = (1, 0)$ (right). The direction vectors are represented by the orange arrows. The values of τ are $\{0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95, 0.99\}$.

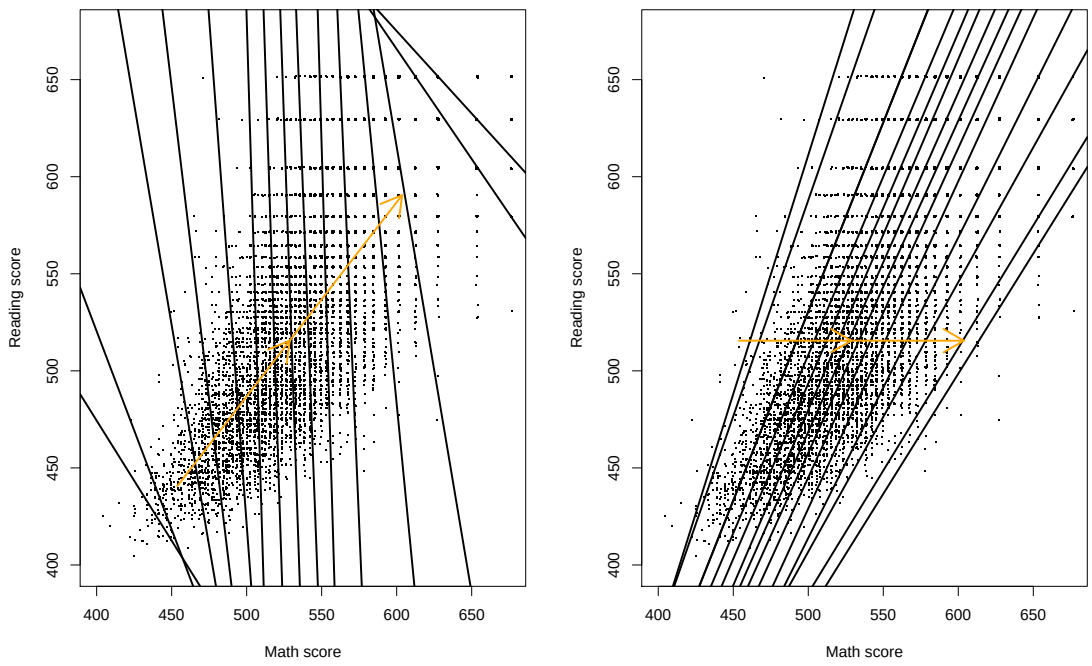

Figure 3.6: Left, fixed $u = (1/\sqrt{2}, 1/\sqrt{2})$ contours. Right, fixed $u = (1, 0)$ contours.

The hyperplanes to the far left of either graph are for $\tau = 0.01$ and as one travels in the direction of the arrow, the next hyperplanes are for larger values of τ, ending with $\tau = 0.99$ hyperplanes on the far right.

The left plot shows the hyperplanes initially bend to the left for $\tau = 0.01$, go nearly vertical for $\tau = 0.5$ and then begin bending to the left again for $\tau = 0.99$. On the other hand the hyperplanes in the right plot are all parallel (roughly). To visualize why this is happening, imagine you are traveling along $\mathbf{u} = (1/\sqrt{2}, 1/\sqrt{2})$ through the Tukey median. Data can be thought of as a viscous liquid that the hyperplane must travel through. When the hyperplane hits a dense region of data, that part of the hyperplane is slowed down as it attempts to travel through it, resulting in the hyperplane tilting towards the region with less dense data. Since the density of the data changes as one travels through the $\mathbf{u} = (1/\sqrt{2}, 1/\sqrt{2})$ direction, the hyperplanes are tilting. However, the density of the data in the $\mathbf{u} = (1, 0)$ direction does not change much, so the tilt of the hyperplanes does not change.

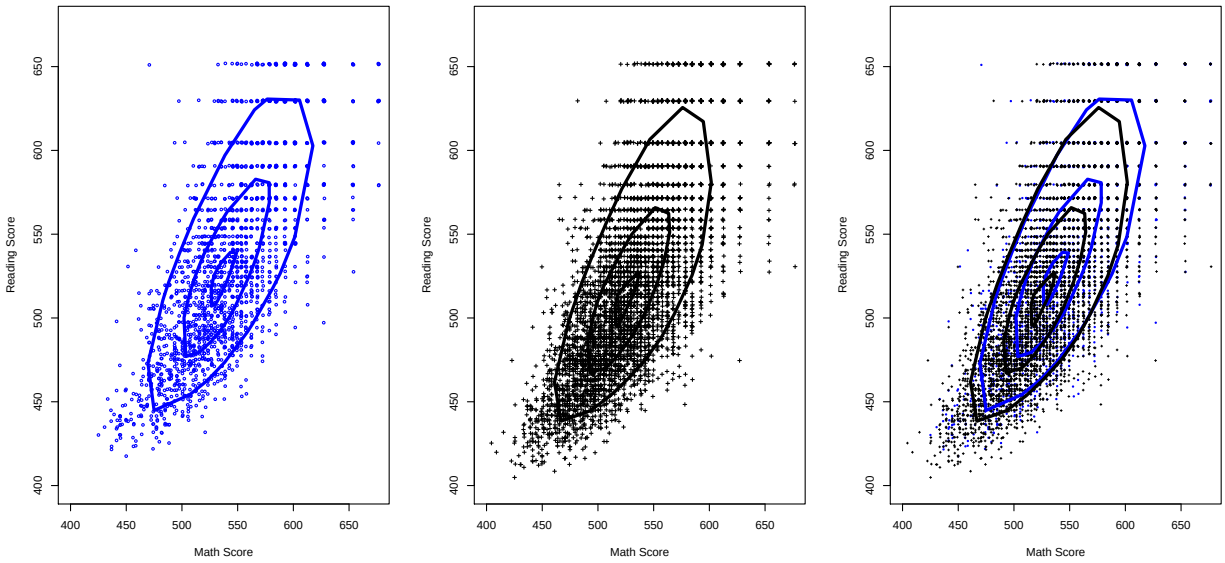

Figure 3.7: Fixed-τ contours. Left, small classrooms. Middle, large classrooms. Right, small and large classrooms overlaid.

Figure 3.7 shows the fixed-τ quantile regions for $\tau = 0.05$, 0.20 and 0.40. The data is stratified into two sets: smaller classrooms (left) and larger classrooms (middle). The quantile regions are overlaid on the third (right) plot. The innermost contour is the $\tau = 0.40$ region, the middle contour is the $\tau = 0.20$ region and the outermost contour is the $\tau = 0.05$ region. Contour regions for larger τ will always be contained in regions of smaller τ. All the points that lie on the contour have a Tukey depth τ of the given contour. The contours for larger τ capture the effects for the more extreme students (e.g. students who perform expceptionally well on math and reading or expceptionally poorly on math but well on reading). The contours for smaller τ capture the effects for the more central or 'median' or 'average' student (e.g. students who do not stand out from their peers). It can be seen that all the contours shift up and to the right for the smaller classroom. This states that the centrality of reading and math scores improves for both for smaller classrooms compared to larger classrooms. Further, this also means all quantile subpopulations of scores improve for students in smaller classrooms.

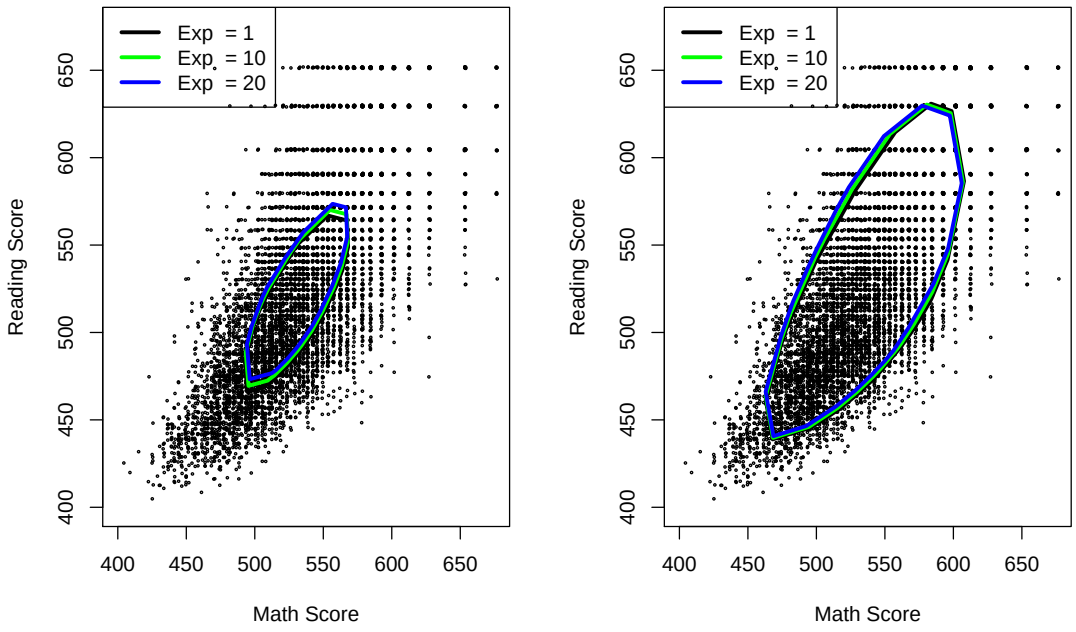

Figure 3.8: Regression tubes (linear). Left, fixed $\tau = 0.2$ regression tube. Right, fixed $\tau = 0.05$ regression tube.

Up to this point only quantile locations have been estimated. When including covariates the fixed-τ regions become 'tubes' that travel through the the covariate space. Since teachers were randomly assigned as well, we can treat teacher experience as exogenous. Then we can estimate the impact of the experience on student outcomes. The new model is

$$\begin{aligned}
\mathbf{Y}_{\mathbf{u}i} &= math_i u_1 + reading_i u_2 \\
\mathbf{Y}^{\perp}_{\mathbf{u}i} &= math_i u^{\perp}_1 + reading_i u^{\perp}_2 \\
\mathbf{X}_i &= years\, of\, teacher\, experience_i \\
\mathbf{Y}_{\mathbf{u}i} &= \alpha_{\boldsymbol{\tau}} + \beta_{\boldsymbol{\tau Y}} \mathbf{Y}^{\perp}_{\mathbf{u}i} + \beta_{\boldsymbol{\tau X}} \mathbf{X}_i + \epsilon_i \\
\epsilon_i &\overset{iid}{\sim} ALD(0, 1, \tau) \\
\theta_{\boldsymbol{\tau}} = (\alpha_{\boldsymbol{\tau}}, \beta_{\boldsymbol{\tau}}) &\sim N(\mu_{\theta_{\boldsymbol{\tau}}}, \Sigma_{\theta_{\boldsymbol{\tau}}}).
\end{aligned}$$

Figure 3.8 shows the fixed-τ quantile regions with a regressor for experience. The values τ

takes on are 0.20 (left plot) and 0.05 (right plot). The tubes are sliced at 1, 10 and 20 years of teaching experience. The left plot shows reading scores increase with teacher experience for the more 'central' students but there does not seem to be a change in mathematics scored. The right plot shows a similar story for most of the 'extreme' students. However, the top right portion of the slices (students who perform best on mathematics and reading) decreases with increasing teacher experience. The best students seem to be performing slightly worse the more experienced a teacher is. A possible story is more experienced teachers try to focus on the class as a whole and tend to focus on the struggling students instead of the high achieving students. The downward shift is small and likely not statistically significant.

Previous research has shown strong evidence that the effect of teacher experience on student achievement is highly non-linear. Specifically the marginal effect of experience tends to be much larger for teachers that are at the beginning of their career than mid-career or late-career teachers (Rice, 2010). We can investigate this non-linearity by adding a quadratic term to the regression equation. The new model is

$$
\begin{aligned}
\mathbf{Y}_{\mathbf{u}i} &= math_i u_1 + reading_i u_2 \\
\mathbf{Y}^{\perp}_{\mathbf{u}i} &= math_i u^{\perp}_1 + reading_i u^{\perp}_2 \\
\mathbf{X}_i &= years\, of\, teacher\, experience_i \\
&\quad + years\, of\, teacher\, experience^2_i \\
\mathbf{Y}_{\mathbf{u}i} &= \alpha_{\boldsymbol{\tau}} + \beta_{\boldsymbol{\tau y}} \mathbf{Y}^{\perp}_{\mathbf{u}i} + \beta_{\boldsymbol{\tau x}} \mathbf{X}_i + \epsilon_i \\
\epsilon_i &\overset{iid}{\sim} ALD(0, 1, \tau) \\
\theta_{\boldsymbol{\tau}} = (\alpha_{\boldsymbol{\tau}}, \beta_{\boldsymbol{\tau}}) &\sim N(\mu_{\theta_{\boldsymbol{\tau}}}, \Sigma_{\theta_{\boldsymbol{\tau}}}).
\end{aligned}
$$

The results are shown in Figure 3.9. It is clear there is a larger marginal impact on student

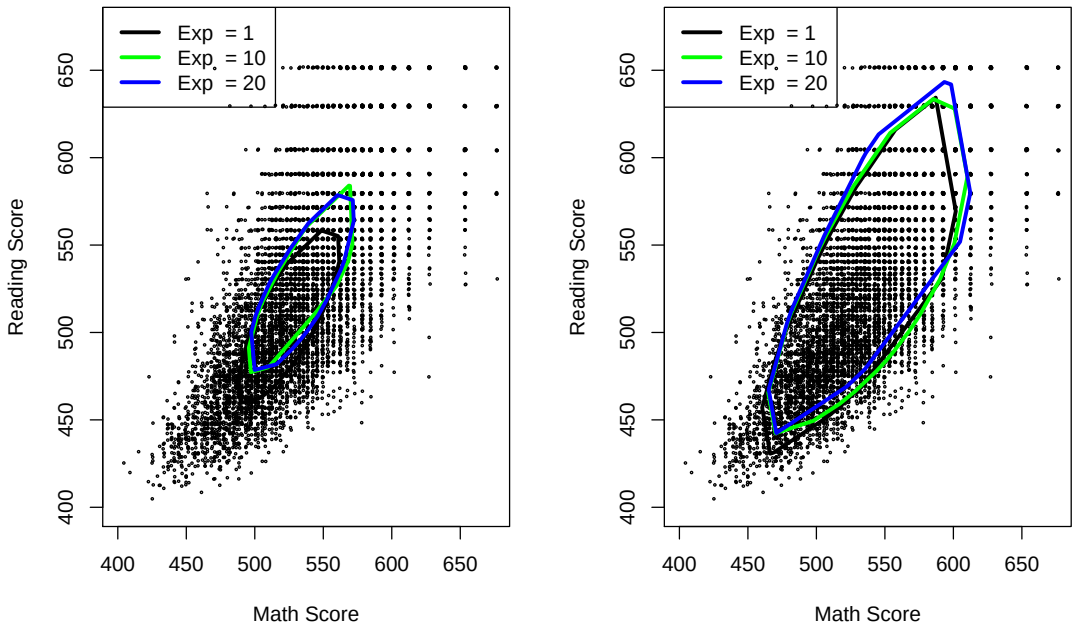

Figure 3.9: Regression tubes (quadratic). Left, fixed $\tau = 0.2$ regression tube. Right, fixed $\tau = 0.05$ regression tube.

outcomes going from 1 to 10 years of experience than from 10 to 20 years of experience. This marginal effect is more pronounced for the more central students ($\tau = 0.2$).

Figure 3.10 shows posterior sensitivity to different prior specifications of the location model with directional vector $\mathbf{u} = (0, 1)$ pointing 90° in the reading direction. The priors are compared against the frequentist estimate (solid black line). The first specification is the (improper) flat prior (i.e. Lebesgue measure) represented by the solid black line and cannot be visually differentiated from the frequentist estimate. The rest of the specifications are proper priors with common mean, $\mu_{\theta_\tau} = \mathbf{0}_2$. The dispersed prior has covariance $\Sigma_{\theta_\tau} = 1000\mathbf{I}_2$ and is represented by the solid black line and cannot be visually differentiated from the frequentist estimate or the estimate from the flat prior. The next three priors have covariance matrices $\Sigma_{\theta_\tau} = diag(1000, \sigma^2)$ with $\sigma^2 = 10^{-3}$ (dashed green), $\sigma^2 = 10^{-4}$ (dotted blue) and $\sigma^2 = 10^{-5}$ (dash dotted red). As the prior becomes more informative β_τ is converging to

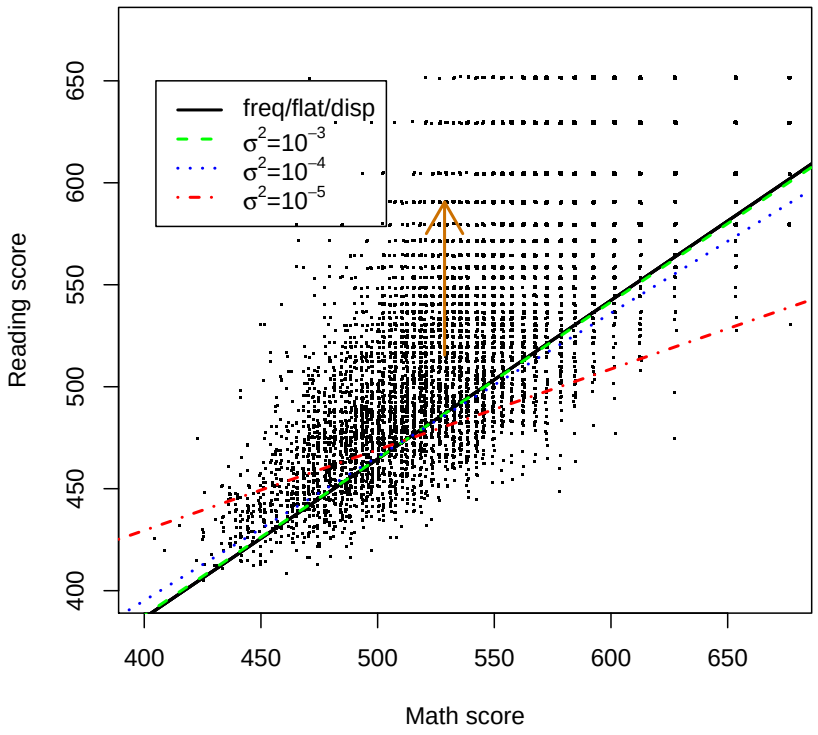

Figure 3.10: Prior influence ex-post

zero with resulting model $\hat{read}_i = \alpha_{\boldsymbol{\tau}}$.

3.6 Conclusion

This paper provides a Bayesian framework for estimation and inference of multiple-output directional quantiles. The resulting posterior is consistent for the parameters of interest, despite having a misspecified likelihood. By performing inferences as a Bayesian one inherits many of the strengths of a Bayesian approach. The model is applied to the Tennessee Project STAR experiment and it concludes that students in a smaller class perform better for every quantile subpopulation than students in a larger class.

A possible avenue for future work is to find a structural economic model whose parameters relate directly to the subgradient conditions. This would give a contextual economic interpretation of the subgradient conditions. Another possibility would be developing a formalized econometric test for the distribution comparison presented in Figure 3.7. This would be a test for the ranking of multivariate distributions based off the directional quantile.

Chapter 4

Strategic Recusals at the United States Supreme Court

4.1 Introduction

In February 2016 Justice Scalia, the leader of the Supreme Court's conservative block, unexpectedly passed away. This vacancy left the court with four liberal justices, three conservative justices and one arguably independent justice (Kennedy). This opened conversation on how outcomes of future cases were going to change without Scalia's presence. Are there going to be more tied outcomes? Are decisions going to be more narrow to prevent ties? Will the threat of a tie pressure justices to switch sides? Are justices going to switch sides with 'under the table' agreements? These last three questions are examples of strategic behavior.

This paper investigates strategic behavior in the justice recusal process. A recusal is when a justice removes himself from a case. A justice is supposed to recuse himself if there is sufficient conflict of interest between himself and the case they are presiding over. Justices at the United States Supreme Court have full power to recuse or not recuse themselves. This

opens the possibility of a justice deciding on a case despite potentially having a conflict of interest. For example, in the 2011 Affordable Care Act case, Justice Thomas and Justice Kagan had the media calling for their recusals because both had a legitimate indication of a conflict of interest, but neither recused. The outcome of the case could have changed if there was a recusal by either one. I expound on this example later.

The Supreme Court has at most nine justices that hear and decide on a case. The decision of the court is given by the decision of the majority of the justices. Thus, if even one justice in the majority has their decision altered by an outside influence, but does not recuse himself, it might effect the decision of the court (especially if it is a 5-4 vote). However, if the justice recuses himself and the court ends with a tied 4-4 vote, then the decision of the lower court is upheld but no precedent is set. This may not have been the decision of the court if there was no conflict of interest. This creates an incentive for a justice to hear a case even though they might have a conflict of interest, this can be rationalized under the guise of the "duty-to-sit" doctrine.[1] See the appendix for a more detailed explanation of the Supreme Court process.

In this paper I construct a simple structural model for strategic recusals that accounts for several possible selection effects. Using this model I provide evidence that justices sometimes fail to recuse themselves when they have a conflict of interest and they might change their vote after a recusal. Next, I calibrate the model and investigate how often justices fail to recuse themselves. I find that roughly 45% to 57% of cases have a justice with a conflict of interest, and 44% to 47% of cases have a justice who remain on the case despite having a conflict of interest. Previous research lack a clear identification strategy and has only been able to provide non-causal evidence for the existence of strategic recusals. No research has attempted to measure the frequency of cases that have a conflict of interest but no recusal.

[1] "Duty to sit" was popularized from a memorandum written by Justice William Rehnquist who refused to recuse himself from a case that would likely have ended in a split vote. The origin of "duty to sit" is to prevent judges and justices from recusing themselves to avoid controversial or burdensome cases. However, it can be abused by a justice to remain on a case where they should have recused (Stempel, 2009).

4.1.1 Recusal

Justices remove themselves from a case if there is a conflict of interest, this is called a recusal.[2] Title 28 section 455 of the United States Code (28 U.S.C. §455) provides guidance on when a justice, judge or magistrate should disqualify themselves from a proceeding. They should disqualify themselves when any of the below are true.

- Their impartiality might be reasonably questioned (waivable, to be explained).
- They have a personal bias or prejudice concerning a party or personal knowledge of disputed evidentiary facts.
- They have involvement in the matter as a material witness, in private practice or as a government employee.
- They had previous professional association with lawyers in the case.
- They expressed an opinion concerning the merits of the case.
- Anyone in their household has a financial interest related to the case.
- Anyone in their household is within 3 degrees of relationship to individuals involved in the case.

The first item is waivable, meaning a justice can still hear and decide on the case as long as there is full disclosure on the record of the basis of disqualification. Justices that recuse themselves are not required to provide reasons for why they recused themselves. However, occasionally, they will voluntarily disclose why they recused themselves. Additionally, if a

[2] Recusal also occurs in courts below the Supreme Court. Parties can request a recusal. If requested and the judge refuses to recuse himself then the party can submit an appeal to a higher court. This appeal can be done while the case is still under review. The higher court then makes a judgment on if recusal is necessary.

justice is requested to recuse himself and refuses to, he is not required to provide reasons why.

Since the Supreme Court has no higher court, each individual justice has full control in deciding to recuse himself or not. This leads to a conflict of interest in resolving conflicts of interest. The Affordable Care Act mentioned in the introduction is a good example. Justice Thomas's spouse was politically active in groups opposing the law and Justice Kagan was holding the position of Solicitor General and could have knowledge of the administration's litigation strategy.[3] Democrats called for recusal of Justice Thomas and Republicans called for recusal of Justice Kagan. Neither recused themselves.

The Affordable Care Act had an 'individual mandate' that required individuals to obtain minimum health insurance or pay a penalty. The Court needed to determine the constitutionality of the individual mandate. The mandate could have been deemed constitutional under the Commerce Clause, the Necessary and Proper Clause or under Congress' taxing power. A 5-4 majority with Justice Thomas decided that the mandate was not constitutional under the Commerce Clause or the Necessary and Proper Clause. This affirmed the lower court's decision and was a win for Republicans. If he would have recused the decision of the lower court would have been affirmed, but no precedent would have been set. Thus allowing the issue to come back to the court. The individual mandate was granted constitutionality under Congress's taxing power in a 5-4 vote with Justice Kagan's vote. This decision reversed the lower court's decision and was a win for Democrats. If Justice Kagan had recused herself and the Court voted in a 4-4, the Court would have affirmed the lower court and concluded that the individual mandate was unconstitutional. Sample (2013) defends the view that it was appropriate for Justices Kagan and Thomas to not recuse themselves in this case.

[3] Solicitor General represents or delegates representation of the federal government before the Supreme Court.

Removing a justice who has a conflict of interest seems to be uncontroversial. A justice is supposed to be impartial. If a justice with a conflict would have changed his legal opinion of a case due to the conflict then he cannot be impartial and should be removed. However, if the justice's opinion is unaffected from a conflict then the conflict has no effect on the outcome of the case (if the justice does recuse himself). Title 28 U.S.C. §455 does make some effort to allow for this second situation by allowing a justice to remain on the case "...in which his impartiality might reasonably be questioned" as long as there is "full disclosure on the record of the basis for disqualification." This is rarely done and recusal usually takes place following the non-waivable portion of 28 U.S.C. §455. If there is no replacement for a justice this could cause concern. By removing a justice, there is a potential to change the decision the court would have had even if the conflict would not have effected the justice's judgement. This possibility puts pressure on justices to not recuse themselves even when there is a conflict of interest.

This raises the question if there should be an authority over the Supreme Court to handle recusal concerns. One potential authority is the United States Congress. Arguments in favor congressional oversight usually stem from the Necessary and Proper clause of the constitution authorizing congress to bring the "Supreme Court into being" (Virelli, 2012). This authority has been used before for congress to determine items such as the Court's term, size, and support offices. There have been bills introduced to change the recusal process but they do not typically get much traction.[4] An argument against congressional oversight states that recusal is a judicial concern and thus the decision belongs to the Court.[5]

[4] In 2009 a house judiciary subcommittee wanted to remove the recusal decision from the justice by allowing each party one automatic disqualification. There was resistance to this proposal since this would lead to judge-shopping (Ingram, 2009). In 2011, a bill, H.R. 862, was introduced (but did not pass). The bill would have made it mandatory for a justice who recuses himself to disclose the reasons for recusal. Additionally, if a justice denies a request for recusal they must provide why they denied the request. Unsuccessful recusal requests would be appealed to a committee of current and retired justices. This was controversial because this could be a violation of the Supreme Court being the highest court. However, there are arguments that reviewing an individual justices recusal decision is not a review of the Court's decision, thus maintaining the Supreme Court as the highest court (Wheeler, 2014).

[5] These debates for or against oversight usually resort to some argument based on judicial ethics. Instead, III (2011) uses constitutionality to argue against oversight.

4.1.2 Strategic behavior

In this section I explore some possible avenues of strategic behavior (not limited to recusals). By strategic behavior, I mean that a justice will follow the rules and professional norms except in circumstances where it is personally suboptimal. For example, strategic behavior could be not recusing oneself in order to influence a case despite having a conflict of interest. Another example would be agreeing with the minority but voting with the majority in order to influence the written opinion.

There has been previous research into justices behaving strategically. Supreme Court Justices are appointed for life or until retirement. Over their tenure they get to know each other well. This added experience allows them to be able to predict how other justices will vote. A justice might be able to anticipate when a case may end in a precarious 5-4 vote or a more coherent 9-0 vote. They may also be able to anticipate if the Court will affirm or reverse a case (Arrington and Brenner, 2004). Thus a majority justice anticipating 5-4 split might be hesitant to recuse himself since he would be the deciding justice. Black and Epstein (2005) and Hume (2014) perform analyses exploring this 'strategic recusal' behavior.[6] They find non-causal evidence for strategic recusals. Black and Epstein (2005) explores the differences in the number of recusals among different justices and natural courts. They find that there is substantial variation in recusals. Hume (2014) performs a logit regression of recusal against sets of variables for statutory, policy and institutional concerns. He finds considerable variation in the number of recusals for different types of cases, tenure of a justice and political leaning of a justice. My paper addresses the same question, but using a structural model.

Black and Epstein (2005) and Hume (2014) also found there was a suspiciously low number of cases that had a recusal and ended in a tie. In other words, if there was a recusal, a

[6] By strategic recusal, I mean not always recusing one's self when one should.

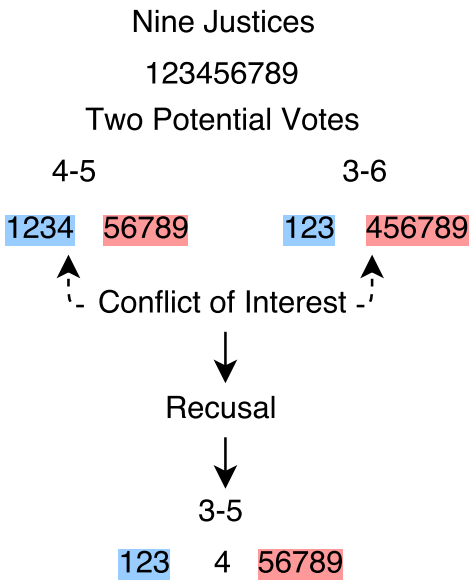

Figure 4.1: 3-5 vote diagram

case was very unlikely to end in a tied 4-4 or 3-3 vote. This could be explained by strategic justice behavior. For example a justice might fail to recuse himself if he anticipates the resulting vote ending in a tie. Alternatively, if a vote is anticipated to end in a tie then the two parties may try to persuade a member of the other party to join their side. A justice might be persuaded by either changing his legal opinion on the case or striking a tit-for-tat deal, voting against his legal opinion. The new member in this (new) majority party will have power to influence the written opinion of the court. If the justice changed his vote to strike a tit-for-tat deal then 3-5 and 5-3 votes might have justices voting against their legal opinion. Thus it is difficult to determine what this vote would have been without recusal or persuasion. For the moment consider no persuasion in a tie, see Figure 4.1. This figure shows how a 3-5 outcome can originate from a 4-5 or 6-3 (without persuasion). There are nine justices (labeled 1,2,3,4,5,6,7,8 and 9) that can vote to affirm or reverse (blue or pink). Justice 4 has a conflict of interest and removes himself from the case, the outcome is a 3-5 vote.

Now consider a variation on a narrative presented in Black and Epstein (2005), suppose that

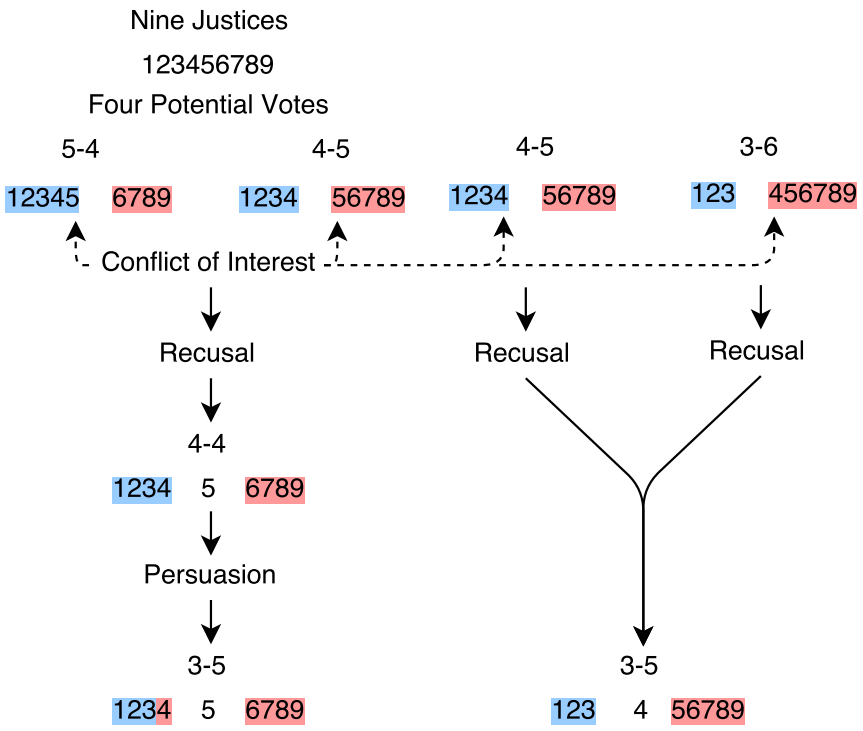

Figure 4.2: 3-5 vote diagram with switching

justices might try to persuade each other to change their vote to prevent a tie. Figure 4.2 depicts this scenario. A vote that ended in a 3-5 could have originated from a 5-4, 4-5, or 6-3. If a vote was going to end in a tie and successful persuasion ensued convincing the marginal (or undecided) justice to join the (new) majority, then the vote could have originated from a 5-4 or a 4-5. Therefore a potential 4-5 vote could result in persuasion or no persuasion depending on who has the conflict of interest. If there was a recusal but no tie, the outcome could have originated from a 4-5 or a 3-6. If a justice did not participate and the observed outcome was a 3-5 then it is hard to tell if the vote was subject to persuasion or not. The story for 5-3 votes is analogous.

If a 3-5 vote was subject to persuasion then I would anticipate it to have a long deliberation reflecting the persuasion process. Deliberation is defined as the number of days between the last oral argument and issuing of the opinion of the court. This duration includes writing of the opinion plus any possible persuasion. Thus a case with no persuasion could still have a

long deliberation if the writing of the opinion takes a long time. Table 4.1 shows the median deliberation for different splits. With one recusal, the shortest deliberation is the 4-4 tie with 29.5 days and the second shortest are the coherent 8-0 and 0-8 votes with 62 days. The longest deliberation with a recusal are the 3-5 and 5-3 votes with 91 days. This provides evidence that 3-5 and 5-3 outcomes were supposed to be 4-4 but were subject to persuasion. The 4-4 vote had the shortest deliberation. This could be due to there being no marginal or undecided justices. The 8-0 and 0-8 votes were second shortest which is likely due to there being no justices in minority to persuade. When all justices participate the longest deliberation is for 4-5 and 5-4 votes with 94 days. As with 4-4, this could be explained by two four-person parties trying to convince the last justice to join their party.

Maj/Min Split (1 recusal)	0	2	4	6	8
Median Deliberation (days)	29.5	91	83.5	84	62
Maj/Min Split (0 recusals)	1	3	5	7	9
Median Deliberation (days)	94	85	83	69	63

Table 4.1: Median deliberation by split

Spriggs et al. (1999) investigates more general contentions in forming a majority on a case. Once the opinion is written it needs a signature of a majority of justices before it becomes the official opinion of the court. Thus the opinion can go through several revisions before it obtains the necessary number of signatures. They find that a given justice in the majority will sign onto the first draft of the opinion 80% of the time. In the other 20% of cases there is some delay in a justice's signing. In 58% of cases there is some form of bargaining that could delay the signing of an opinion. Additionally, a justice's decision to influence the opinion of the majority is not based entirely off ideological distance from the writer, factors such as size of coalition and previous interactions with the author of the opinion play a role as well.

There are other kinds of strategic behaviors at the Supreme Court. Since justices vote in descending order of seniority, junior justices could be influenced by senior justices. This invites a possibility for junior justices to change their opinion and switch their vote to side

with senior justices after seeing their vote and hearing their reasoning. However, it is difficult to determine if a justice switched their vote with the intent of siding with senior justices or because they were persuaded by the arguments made by the senior justices. Arrington and Brenner (2004) find that switching does not occur often. Since justices provide reasoning for their decision during the voting process, if a justice is unsure on his decision he can 'pass' and vote at the end, after hearing the others opinions. This feature of the voting process can be abused by passing with the intention to vote with the majority. By being in the majority the justice can then influence the opinion of the court. Johnson et al. (2005) find that Chief Justice Burger, Chief Justice Rehnquist, Justice Douglas, and Justice Brennan used their ability to pass in order to vote with the majority and influence the opinion. However, Arrington and Brenner (2004) conclude it is rare. Even though all justices in the majority contribute to the opinion, it is written by one justice. The writing of the opinion is assigned by the senior justice in the majority. The senior justice in the majority may strategically assign the opinion to the justice most closely aligned in ideology with himself. Wahlbeck (2006) finds evidence of this in the Rehnquist court from terms 1986 to 1993.

The take away from this section is that there are many ways justices could be behaving strategically and there is evidence to support it. Thus there it is suspect that justices behave strategically with respect to the recusal process as well.

4.2 Data

The majority of the data used in this analysis comes from the Supreme Court Database[7] and the rest is from the U.S. Supreme Court Justices Database.[8]

The model presented in section 3 requires vote-level data for two variables to show that justices recuse themselves strategically. The first variable is an indicator for when a justice recuses himself and the second is the total number of votes affirming for a given case. The Supreme Court Database only records when a justice did not vote, it does not record why the justice did not vote. Usually a justice misses a vote due to sickness, recusal, or being appointed to or leaving the court midterm. Following Black and Epstein (2005) and Hume (2014) I classify a missing vote as sick when a justice misses all oral arguments for at least 4 cases for 2 or more consecutive days of oral argument. Appointment and withdrawal dates are available in the U.S. Supreme Court Justices Database. All votes missed that were not due to sickness, midterm appointment, or midterm withdrawal were labeled as recusals.[9]

4.2.1 Data exploration

On a given case presented to the Supreme Court there may be multiple issues that have to be voted on. These are called 'case issues,' but I will simply call them 'cases' for ease of reading. The justice level vote data on cases dates back to 1946. Since then there been

[7] The database contains vote level data on Supreme Court case issues dating back to 1946 and is updated annually. It is free and publicly available at http://supremecourtdatabase.org. A Supreme Court case may contain multiple issues that need to be voted on separately, these are called case issues. The analysis done in this paper is done with respect to votes on given case issues. For ease of readability I will referred to them just as cases instead of case issues.

[8] The database contains demographic, biographical, and professional data on Supreme Court Justices. The data is free and publically available at http://epstein.wustl.edu/research/justicesdata.html.

[9] This method of classifying missing votes does not classify recusals with perfect accuracy. For example, if a justice were to recuse himself from all cases totaling 4 or more over 2 or more consecutive days of oral arguments, then those missing votes would be classified as sickness and not recusal. Alternatively, if a justice misses 2 votes from one day of oral arguments from being sick and votes on cases from immediately previous and future arguments then those votes will be labeled as recusals and not sickness.

12,907 different cases with a total of 113,401 votes. Of those 3,323 instances where a justice did not vote and 2,010 were due to recusal. Of the recusals 1,680 cases had 1 justice recuse himself, 302 cases had 2 justices recuse themselves, 27 cases had 3 justices recuse themselves and 1 case had 4 justices recuse themselves. There were no cases with more than 4 justices recusing themselves.

If a case had a recusal and resulted in a tie, the recusal made a difference in the outcome.[10] There have been total 12,907 cases, and 97 had ended in a tie vote (0.75%). Thus the decision of the lower court in those cases was upheld, no opinion was written and no precedence was established. Of the 97 ties, 71 cases had at least one recusal. If there were multiple recusals in a case then the decisive outcome of a case have flipped (e.g. a 5-4 turning into a 3-4 after recusal). There 53 case issues with this potential outcome. Thus, if there was no strategic behavior, a total of 124 cases could have had a different outcome if justices did not recuse themselves.[11]

Recusal rates among justices is highly heterogeneous. Some will recuse themselves quite often and some very rarely. See Figure 4.3 for a comparison of recusals by justice. The x-axis of both plots is the first (and second) initial followed by the last name of a justice. The top bar plot shows the total number of recusals a justice makes over their entire tenure at the court. The justices with the most recusals are Thurgood Marshall (328 recusals), William O. Douglass (260 recusals), and Lewis Powell (250 recusals) and the justices with the least number of recusals are Ruth Bader Ginsburg (2 recusals), Earl Warren (5 recusals)and Charles Evans Whittaker (5 recusals).[12] The bottom bar plot shows the percentage of recused

[10] If the court would have voted to reverse the lower court, the recusal clearly made a difference. If the court would have voted to affirm the lower court then the vote would not have made a difference but the opinion could have changed policy outcomes.

[11] 124 = ties + flips.

[12] Some recusals are more procedural than others. For example, if a justice served as Solictor General or as a judge on a lower court before being appointed to the supreme court, they will recuse themselves for any cases they were apart of on their previous appointment. Thus it is quite common for justices to have a large number of recusals early in their tenure at the Court. For example, Justice Thurgood Marshal served as Solicitor General prior to his Supreme Court appointment. Over his 23 year tenure at the Supreme Court he had 328 recusals. However, 152 recusals were during his first year (46% of his total). In his first 3 years

votes over the total number of votes a justice made. The justices with the largest percentage of recusals are Elena Kagan (11.4%), Abe Fortas (10.4%) and Robert H Jackson (9.8%) and the justices with the smallest percentage of recusals are Ruth Bader Ginsburg (0.08%), Earl Warren (0.15%), and Potter Stewart (0.20%). William J. Brennan had 8,041 total votes and no recusals over his entire career.

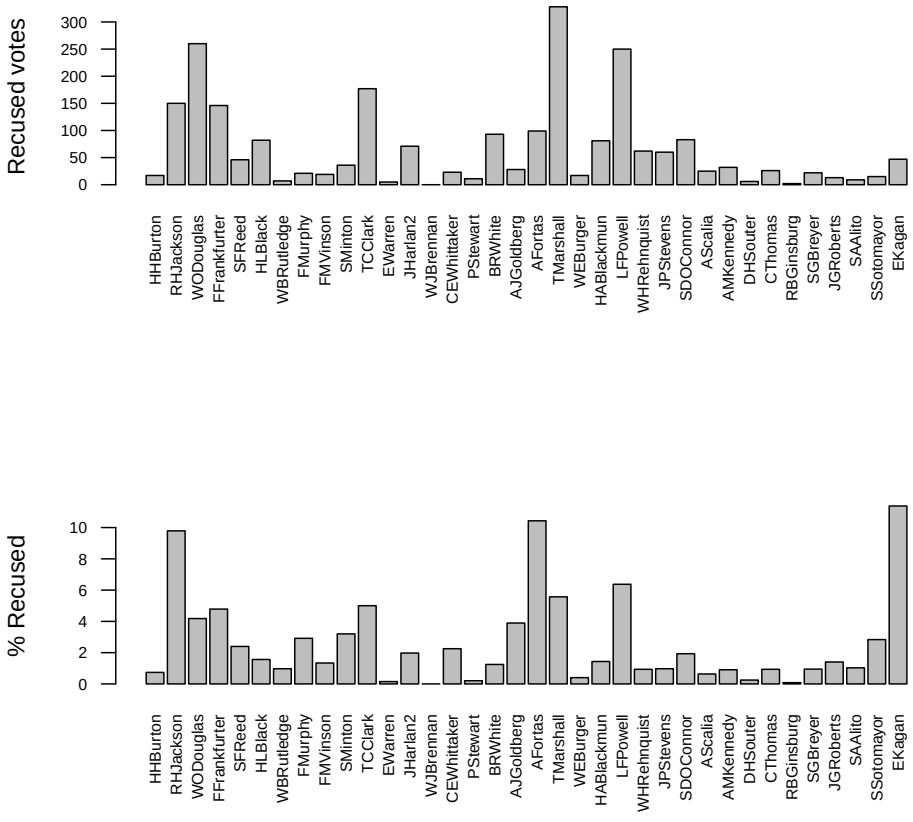

Figure 4.3: Top, number of recused votes over tenure per justice. Bottom, percentage of recused votes per justice.

There is also much heterogeneity in the number of recusals over time. See Figure 4.4 for a comparison of recusals by court term. The solid black line shows number of votes that were recused during the given term and the gray dotted line shows the percentage of votes

he had 217 recusals (66% of total). For the last 20 years of his tenure he had 111 recusals, which was 34% of his total.

that were recused during a given term. It can be seen that recusals were more common before 1988. This is because the Court heard more cases (annually) before 1988 than after.[13] Another observation is the total number recused and percentage recused move proportionally to each other. If the percent of cases with a conflict of interest was constant, then the percentage of recusals should be flat.[14] This percentage could be changing due to changes in the type of cases the court hears, which justices are on the court, or the philosophy on recusals (e.g. not granting cert to cases where there might be recusals).

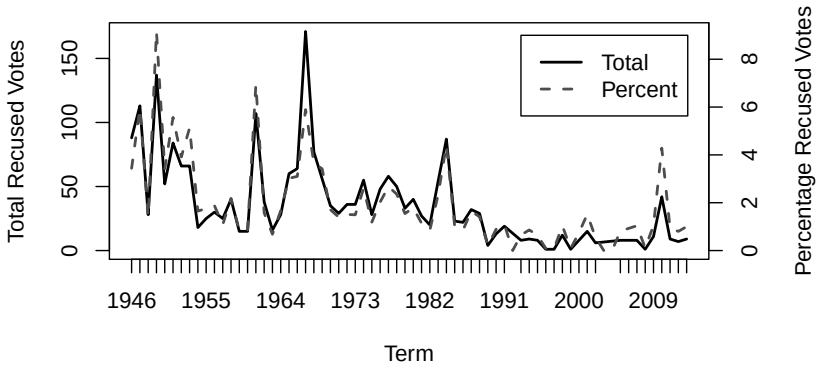

Figure 4.4: Recused votes per term

4.3 Recusal decision model

In this section I develop the model used to investigate the existence of strategic recusals. My strategy is to show that the difference between votes to affirm and votes to reverse is independent of recusal if justices always recuse themselves when they have a conflict of

[13] From 1946 to 1988 the court heard between 150-300 cases a year. Starting in 1988 till mid 2000's the court tended to hear less cases each year and eventually settled on hearing about 100 cases a year, which continues till 2014.

[14] This can formally be tested. The hypothesis is $H_0 : \frac{\#recusals}{\#votes} = c$ for some constant c. This can be performed by regressing $log(number\ recused\ votes_t) = \beta_0 + \beta_1 log(number\ total\ votes_t)$ and testing for significance of $H_0 : \beta_1 = 1$ vs $H_0 : \beta_1 \neq 1$. Which is strongly rejected (p-value $= 3.12 \times 10^{-3}$).

interest. I then show that independence does not hold and argue that it is due to justices recusing themselves strategically. I then provide robustness checks. Proofs and some lemmas are in the appendix.

I assume that justices have a definite opinion on a case and they have no uncertainty in their own opinion.[15] Thus for a fixed justice I assume a justice knows his own vote but other justices' votes are random from his perspective and all votes are random to the researcher. For a given case, let X be the number of votes affirming and let Y be number of votes reversing. X and Y are random to the researcher. For now assume all 9 justices vote. Notice that the decision of the court is determined by $X - Y$. Where $X - Y \geq 0$ means the court affirms and $X - Y < 0$ means the court reverses. Also note that $Y = 9 - X$ implies $X - Y = 2X - 9$. Thus the decision of the court is uniquely identified by knowing X.

Let c_i be an indicator for when justice i has a conflict of interest.[16] Let C_X be the number of justices with a conflict of interest affirming and let C_Y be the number of justices with a conflict of interest reversing. The support of C_X is $\{0, 1, ..., X\}$, likewise the support for C_Y is $\{0, 1, ..., 9 - X\}$ and the support for total conflicts of interest, $C_X + C_Y = \sum_{i=1}^{9} c_i$, is $\{0, 1, ..., 9\}$. The first assumption states that in a given case each justice has some random probability of having a conflict of interest (allowing for heterogeneity across justices and cases).

Assumption 4.1. *Suppose* $X \overset{iid}{\sim} Multinomial(1, q)$ *where* $q = (q_0, q_1, ..., q_9)$ *and* $c_i|p_i \overset{\perp}{\sim} Bernoulli(p_i)$ *for* $i \in \{1, 2, ..., 9\}$ *where* $p_i \overset{\perp}{\sim} F_{p_i}$*, and* $E[p_i] = \mu_{p_i}$ *for* $i \in \{1, 2, ..., 9\}$*. Where* F_{p_i} *is some unknown CDF.*

This assumption is fairly general and I would not anticipate it generating much controversy. Assumption 4.1 will need to be restricted for a lemma required later on. Assumption 1.1b

[15] If this assumption is violated then effects of justices changing their opinion and justices voting against their opinion are confounded.

[16] It is not necessary to define what exactly a conflict of interest is, but it is fine to define it according to 28 U.S.C. §455 for the context of this paper.

is this restriction. It says that each justice has a random probability of conflict of interest and common mean probability, μ_p.

Assumption 1.1b. $\mu_{p_i} = \mu_{p_j} = \mu_p$ *for* $i, j \in \{1, 2, ..., 9\}$

Assumption 1.1b does not allow for heterogeneity in mean of the probability of conflict of interest over justices. I will defer further discussion of Assumption 1.1b till later. I can now present the main theorem.

Theorem 4.1. *Suppose Assumption 4.1 holds then* $C_X + C_Y \perp X$. *Thus* $C_X + C_Y | X \sim C_X + C_Y$. *This result still holds for Assumption 1b as well.*

This theorem is of interest because it gives a testable implication, $C_X + C_Y \perp X$ (to be tested using a χ^2 test of independence). However, $C_X + C_Y$ might not be observed. Additionally, if there is a recusal then X might not be observed because $X - Y$ is not observed due to X and Y counting the votes for when all justices vote. The next assumption is assumed for the sake of contradiction and allows us to observe $C_X + C_Y$.

Assumption 4.2. *If a justice has a conflict of interest he will recuse himself.*

Under Assumption 4.2, $C_X + C_Y$ is observed where the number of conflicts of interest is equal to the number of recusals (which is observed). Define $X^\star = X - C_X$ and $Y^\star = Y - C_Y$ to be the observed votes under Assumption 4.2. Thus $X = X^\star + C_X$ is potentially unobserved under Assumption 4.2. This is problematic for testing Theorem 4.1 since I need to observe X when $C_X + C_Y = 0$ and $C_X + C_Y = 1$ (it is sufficient to consider only these cases and ignore $C_X + C_Y > 1$). Given $C_X + C_Y = 1$, X could take on two different values ($X = X^\star$ or $X = X^\star + 1$). Using this, I can impute X when $C_X + C_Y = 1$. Given the observed $X^\star$, I can count each potential vote (i.e. $X^\star$ and $X^\star + 1$) as an observation and weigh them appropriately (i.e. count $w_0 * X^\star$ and $w_1 * (X^\star + 1)$ with weights w_0 and w_1). The weights can be equated to the probability for each potential vote. That is $w_0 = Pr(C_X = 0 | C_X + C_Y = 1)$

and $w_1 = Pr(C_X = 1|C_X + C_Y = 1)$. The next lemma provides a distribution used to calculate the weights.[17]

Lemma 4.1. *Suppose Assumptions 4.1, 1.1b and 4.2 hold then* $Pr(C_X = 0|X, C_X + C_Y = 1) = 1 - \frac{X}{9}$ *and* $Pr(C_X = 1|X, C_X + C_Y = 1) = \frac{X}{9}$.

Since X is not observed we take its expectation, $E[X|X^\star]$. Conditioning on $X^\star$ is just a truncation. The support of $X|X^\star$ becomes $\{X^\star, X^\star + 1\}$.

Lemma 4.2. *Suppose Assumptions 4.1, 1.1b and 4.2 hold,* $C_X + C_Y = 1$ *then* $E[X|X^\star] = X^\star + \frac{q_{X^\star+1}}{q_{X^\star}+q_{X^\star+1}}$.

Thus the weights are $w_0 = 1 - \frac{X^\star + \frac{q_{X^\star+1}}{q_{X^\star}+q_{X^\star+1}}}{9}$ and $w_1 = \frac{X^\star + \frac{q_{X^\star+1}}{q_{X^\star}+q_{X^\star+1}}}{9}$. The parameters $q_{X^\star}$ and $q_{X^\star+1}$ can be replaced with their sample estimates.[18]

Lemmas 4.1 and 4.2 used Assumptions 4.1 and 1.1b which restricted the mean probability of conflict of interest to be homogeneous over justices. This implies that for a given case each justice will have the same *average* probability of having a conflict of interest. It's clear from Figure 4.3 that justices recuse themselves at different rates. However, the key word is *average*, the 'realized' probabilities are free to vary among justices. Assumption 1b is used to simplify the distribution of $C_X|C_X + C_Y$ which was used to calculate the weights w_0 and w_1. Without it the weights would have to be calculated on observation-by-observation basis and would likely not change the resulting aggregate counts much.

Using Lemma 4.2 the weighted votes can be imputed and the contingency table is presented in Table 4.2. The columns are the number of justices who voted to affirm, this ranges from

[17] Notice that this lemma requires Assumption 1.1b, the strengthened version on Assumption 4.1. Without Assumption 1.1b, I would have to impose restrictions on the distribution of probabilities of conflict for individual justices, F_{p_i}. Additionally, the resulting distribution of $C_X|X, C_X+C_Y$ would be *very* complicated.

[18] The maximum likelihood estimate is just the sample proportion for each X value for the cases where all 9 justices vote. The estimate is $\hat{q} = (0.241, 0.070, 0.091, 0.109, 0.114, 0.103, 0.073, 0.051, 0.039, 0.110)$. There could be some contamination from strategic behaviors but the resulting inferences are robust to the choice of weight. The weights $w_0 = 1 - \frac{X^\star}{8}$ and $w_1 = \frac{X^\star}{8}$ result in the same inferences.

0 to 9. The 'Recusal' row is the number of votes to affirm when there was a recusal. Notice that there were no outcomes with 9 votes to affirm, this is since only 8 justices participated. The 'Recusal Imputed' row is the (rounded) imputed counts from the row above it (e.g. there was an imputed count of 129 outcomes for when there was 3 votes to affirm with a recusal, $129 = round\left(\left(1 - \frac{3+\frac{0.114}{0.109+0.114}}{9}\right)144 + \left(\frac{2+\frac{0.109}{0.091+0.109}}{9}\right)147\right)$). The 'No Recusal' row are the counts to affirm when all 9 justices participated.

	0	1	2	3	4	5	6	7	8	9
Recusal	344	150	147	144	67	80	98	90	194	0
Recusal Imputed	335	132	132	129	90	65	76	85	80	188
No Recusal	2158	623	817	974	1024	919	657	454	348	990

Table 4.2: Contingency table of votes

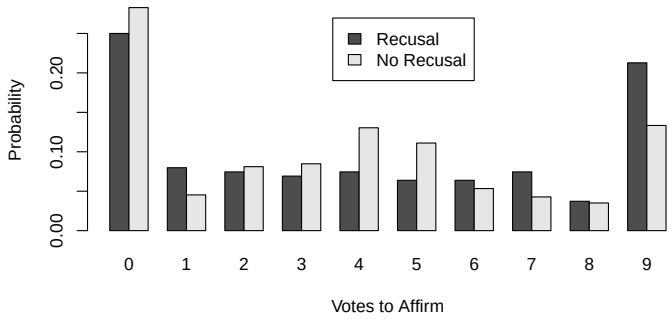

Figure 4.5: Distribution of affirmations stratified by recusal (with imputation)

The bottom two rows are used to test Theorem 4.1 (using a χ^2 test of independence $H_0 : C_X + C_Y \perp X$ vs $H_a : C_X + C_Y \not\perp X$). The resulting test rejects the null hypothesis ($p < 2.2 \times 10^{-16}$) and concludes $C_X + C_Y \not\perp X$. Rejection of the null hypothesis is unsurprising. By inspecting barplots of the relative frequencies (see Figure 4.5 using imputing and Figure 4.6 not using imputing), it is clear that they are not independent. If they were independent, each pair of bars would be approximately the same height. Thus, either (1) Theorem 4.1 does not hold and thus Assumption 4.1 is violated or (2) the test did not test Theorem 4.1.[19] These concerns can be addressed with robustness checks (presented below). I first address (2) then

[19] By (2) I mean $H_0 : C_X + C_Y \perp X$ vs $H_a : C_X + C_Y \not\perp X$ was not tested because what really was tested was $H_0 : 9 - X^\star + Y^\star \perp X^{imputed}$ vs $H_a : 9 - X^\star + Y^\star \not\perp X^{imputed}$. If $9 - X^\star + Y^\star \not\perp X^{imputed}$ does not imply $C_X + C_Y \not\perp X$, then Theorem 4.1 was not tested.

(1). The results from the robustness checks are organized in Table (4.3).

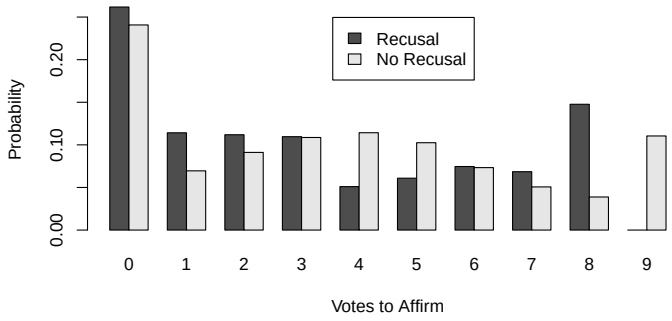

Figure 4.6: Distribution of affirmations stratified by recusal (without imputation)

If (2), the χ^2 test did not test Theorem 4.1, there could be several reasons for this.[20] Consider a fixed case i, the realized vote could change for different counterfactuals of conflicts of interest. Mathematically, $(X = x_i|C_X + C_Y = c_1) \neq (X = x_i|C_X + C_Y = c_2)$ where $c_1 \neq c_2$, meaning that the number of (possibly unobserved) votes to affirm changes with the counterfactual number of conflicts. Alternatively the population of cases is heterogeneous with respect to conflicts of interest, $Pr(X|C_X + C_Y = c_1) \neq Pr(X|C_X + C_Y = c_2)$ where $c_1 \neq c_2$.

There is a dynamic aspect to the recusal and voting process, they do not occur at the same time. This leads to the possibility that for a given case, X can change with the counterfactual number of conflicts (I am thinking 0 to 1 conflicts). This model is implicitly assuming that for any given case the number of votes affirming, X, is fixed whether it is observed or not. However, this can be a controversial assumption. A simple counterexample is to suppose there was a recusal and the vote ended in a 3-5 due to persuasion. The persuasion process changed the outcome (which would have been 4-4), this was directly due to the recusal (and

[20]This could be the case if Assumption 4.2 was violated. Then the number of recusals would not equal number of conflicts of interest. This is the argument I am making, but this is not the only reason why the test may have rejected the null hypothesis.

One of the simpler alternative explanations could be the algorithm to determine recusals was incorrect. Any misclassification of the algorithm would be random enough that it would not affect the test. For example, it would affect the test if misclassification was not independent of votes to affirm, but this is ridiculous.

The second simpler explanation is a greater concern. The imputing of the vote did not capture the distribution of the true (unobserved) X. The imputation could have been incorrect by using the wrong weights. I re-ran the test using a grid of weights, $w_1 \in \{0, 0.1, 0.2, ..., 1.0\}$ and $w_2 = 1 - w_1$, the test overwhelmingly rejected each time. However, there are numerous other potential weighing schemes.

hence the conflict of interest). This counterexample would mean that votes imputed for $X = 3$ or 5 when there was a recusal did not reflect the 'true' votes. The first counterexample can be avoided by combining the values of X for 3,4,5 and 6 into one category and rerunning the test, which results in a p-value less than 2.2×10^{-16} rejecting the null hypothesis. The problem with combining $X = 3, 4, 5, 6$ to avoid the first counterexample is these are the situations when a justice would be more likely to fail to recuse himself despite having a conflict of interest. This is because his vote is more likely to make a difference in the outcome in these instances. Thus this specification can be thought of as a conservative conclusion. However, by not combining these values the test is effectively testing for strategic recusal or vote switching.

Another counterexample is if X was going to be 1 or 8, leaving 1 justice in the minority, then that justice might switch his vote to align with the majority and show solidarity with the court (Epstein et al., 2013).[21] This counterexample would mean that 0 or 8 votes to affirm with a recusal and 9 votes to affirm without a recusal might not reflect the 'true' votes. The second counterexample can be avoided by combining $X = 0$,1 and 8, 9 and rerunning the test, resulting in a p-value less than 2.2×10^{-16} . This test can be interpreted as a test for strategic recusal or vote switching controlling for solidarity. By combining 0,1 and 3,4,5,6 and 8,9 both counterexamples can be avoided and the resulting p-value is less than 2.2×10^{-16}, rejecting the null hypothesis.

If (1) then Assumption 4.1 is violated and either the either conflicts of interest are not independent between justices, the distribution of the probabilities of conflict is causally related to X or the votes are not independently and identically distributed. Depending on the mechanism of dependence, lack of independence of conflicts between justices would not necessarily invalidate Theorem 4.1. For example, if two justices have large investment portfolios that are very similar then they would have correlated conflicts of interest but this

[21] Notice that this is unrelated with recusals or conflicts of interest. It is a measurement error that would affect the vote whether or not there was a recusal.

would not necessarily cause $C_X + C_Y \not\perp X$. The other option is for $p_i = p_i(X)$ meaning that the probability of a conflict of interest is (partially) determined by the number of justices who vote to affirm. While there might be a possibility of a spurious correlation between p_i and X, I am unsure why there might be a causal connection.

If (1) is true due to the distribution of votes not being independent and identically distributed this test might not be testing Theorem 4.1 is due to a heterogeneity in the population. Specifically, the distribution of the votes to affirm for cases without a conflict of interest is inherently different than those cases with a conflict of interest. Three potential (non-exhaustive) sources of this heterogeneity are the composition of the court, term of the court, and the issue a case pertains to. Term and issue can be separately controlled for, however there is not enough data to effectively control for composition of the court.[22] The analysis used cases dating back to 1946, but looking at Figure 4.4 it appears there were much less recusals after 1988. While the reason for this is unclear, there was some sort of structural change in the Court causing a reduction in recusals.[23] To focus on the more modern court I can restrict the counts to cases after 1988. The resulting p-value is 8.47×10^{-4}. See Table 4.3 for p-values with restrictions on the support of X as well.

In addition to time, another source of heterogeneity is the type of issue a case is addressing. There are 14 categories of issues that a case could pertain to. I will restrict myself to ones that have counts of 1,000 or more cases. This leaves five categories, 1 Criminal Procedure, 2 Civil Rights, 3 First Amendment, 4 Economic Activity, and 5 Judicial Power. Restricting myself to each one of these categories the resulting p-values can be seen in Table 4.3. The columns represent the category that the counts are restricted to. The rows represent if there

[22] The court rarely goes longer than a couple years without there being a change in the justices that comprise the court. There is one uninterrupted 11 year frame where there was no change, this was from terms 1994 to 2004. By restricting the counts to this frame I would wish to test the hypothesis controlling for changes in the composition of the court. However the resulting contingency tables do not satisfy the finite sample conditions.

[23] One possible explanation is the Court became less willing to grant cert to cases where there was a conflict of interest. Another explanation is that justices became less likely to recuse themselves when they had a conflict of interest.

was any support restrictions. The value in each cell is the p-value from the χ^2 test. A cell has a '·' if a finite sample condition was violated and the test was not performed.[24] An interesting (and somewhat ironic) result was that category 5, Judicial Power, has the largest p-values showing the least evidence for strategic recusals.

	Count Restriction						
Support Restriction	None	Post 1988	1	2	3	4	5
None	$< 2.20e$-16	$8.47e$-4	$6.25e$-6	$2.10e$-4	$2.35e$-5	$7.05e$-7	0.364
3,4,5,6 combined	$< 2.20e$-16	$3.91e$-4	$2.91e$-6	$2.22e$-5	·	$1.64e$-7	0.549
0,1 and 8,9 combined	$< 2.20e$-16	$4.51e$-3	$1.08e$-6	$6.28e$-4	$8.19e$-5	$7.15e$-4	0.353
Both Restrictions	$< 2.20e$-16	$2.01e$-3	$3.24e$-7	$5.27e$-5	$2.30e$-5	$2.24e$-4	0.574

Table 4.3: P-values with robustness checks

The conclusions from these individual tests cannot be held simultaneously without a correction. This is because when performing multiple hypothesis tests the type 1 error rate increases with the number of tests. A correction for this is the Bonferroni correction, which is a conservative correction.[25] If a researcher wishes to test at a certain alpha level, the correction rejects if the alpha level divided by the number of tests is less than the p-value. Since 27 tests were performed, the p-value is checked against the alpha level divided by 27 (e.g. an alpha level of .05 should adjust to $.05/27 = 1.85e$-3 and .01 to $.01/27 = 3.70e$-4). Thus most tests reject at the .05 or .01 levels (with correction) and due to the conservative nature of the correction, it is likely that all tests reject (except for type 5 cases).

[24] The condition for a 2×2 table is all expected counts are greater than 10. For tables larger than 2×2, 80% of the expected counts must be greater than 5 and all must be greater than 1. If the condition was violated then the test was not performed. The Yate's correction was not used.

[25] In this situation it is very conservative. The Bonferroni correction provides accurate type 1 errors when the tests are independent, meaning that the result of one test provides no information about the result of another test. The correction provides conservative type 1 errors when there is some dependence in the tests (which is the case here). There is shared observations between the tests performed, so there is some dependence.

4.3.1 Simulation study

In the previous section I showed justices might be recusing themselves strategically. This invites the question, "how often are justices not recusing themselves when they should?" This is not an easy question to answer since it requires observing two unobservables; the probability of conflict of interest (denoted ϕ_1) and probability of recusal given conflict of interest (denoted ϕ_2). I obtain an approximate solution by simulation of a calibrated model. This model considers only one mechanism (strategic recusals) to generate predictions. As discussed in the previous section, this might not be the only mechanism generating the observed results. Thus conclusions drawn from the simulation should only be considered rough estimates.

Define $R \equiv 9 - (X^\star + Y^\star)$ to be the number of recusals. In the last section I showed $R \not\perp X$. This is equivalent to $X|R = r \not\sim X$. Using Kullback-Leibler divergence I can measure 'how violated' the independence condition (i.e. $R \perp X$) is by looking at how 'different' $X|R = r$ and X are.[26] I'll only focus on the case with no recusals, $r = 0$, for reasons that will be clear later. The difficulty from comparing $X|R = r$ with X is the unconditional X is not observed (even when $r = 0$ we observe $X|R = 0$). I derive conditions where $X|R = 1$ can be used as a proxy for X. I then simulate strategic recusals by generating a random variable W with structural parameters ϕ_1 and ϕ_2 to mimic $X|R = 0$. I keep the values of ϕ_1 and ϕ_2 that generate the same Kullback-Leibler divergence of W from $X|R = 1$ equivalent to the divergence of X from $X|R$ (Usine $X|r = 0$ and $Z|r = 1$).

The Kullback-Leibler divergence is a function used to measure how different two distributions are. Its foundations are in information theory and it is a common measure used in the study

[26] There are many ways other than Kullback-Leibler divergence to quantify how dissimilar two random variables are (e.g. Hellinger Distance, Total Variation Distance, etc.). I choose Kullback-Leibler because it measures the amount of information lost when random variable B approximates random variable A. Loosely, it measures the 'distance' between A and B in terms of A. This lets A be act as a reference distribution, in this case A is the marginal X and B is $X|R$.

of misspecified models (**?**). Define A and B to be discrete random variables where A is dominated by B.[27] The Kullback-Leibler divergence of B from A is

$$KL(A;B) = \sum_a Pr(A=a) log\left(\frac{Pr(A=a)}{Pr(B=a)}\right)$$

Note that the Kullback-Leibler divergence is non-symmetric and non-negative. If two random variables have the same distribution almost everywhere, their Kullback-Leibler divergence will be 0. The larger the Kullback-Leibler divergence is, the 'further apart' the two distributions are. Also note that $KL(A;B) = KL(A;C)$ does not necessarily imply $B \sim C$, this is a desirable feature that I will elaborate on later.

Define the estimated Kullback-Leibler divergence to be $\hat{KL}(A;B)$, where $Pr(\cdot)$ is replaced with its relative frequency counterpart $\hat{Pr}(\cdot)$. It is clear that if A and B are discrete and A is dominated by B then $\hat{KL}(A;B)$ is strongly consistent for $KL(A;B)$.

As previously mentioned, the distribution of X is never directly observed, we only observe $X|R = r$. Even when there are no recusals, $X|R = 0$ is observed but the marginal X is not (remember $X|R \not\sim X$). How can I measure the Kullback-Leibler divergence of $X|R = 0$ from X when X is unobserved? The next two assumptions are used to find an observable random variable with the same distribution as X.

Assumption 4.3. *If a justice recuses himself, then there was a conflict of interest.*

This assumption is to prevent the situation where a justice recuses himself for things other than a conflict of interest (e.g. to avoid controversial cases). This assumption is the converse of Assumption 4.2 and is not very controversial for the Supreme Court.[28] The next

[27] Meaning $Pr(B=a) = 0$ implies $Pr(A=a) = 0$.

[28] This assumption would be more controversial at lower courts. A major reason why a judge might recuse himself despite not having a conflict of interest is because the case is controversial (Stempel, 2009). If the case is controversial then the judge would get media attention and this attention might hurt his chances of getting reappointed or reelected after his term is up. The idea of 'duty to sit' is to prevent this. However,

assumption is necessary for the proof of the main theorem in this section. This assumption states that there is no more than one conflict of interest.

Assumption 4.4. $C_X + C_Y \in \{0, 1\}$

This assumption is very restrictive. It is clear Assumptions 4.3 and 4.4 cannot both hold when there are 2 or more recusals in a case (which occurs in about 16% of cases where there is recusal). This appears to be damning, but it can be avoided by using data after 1988 where there was rarely more than 1 recusal in any given case (a robustness check used later).[29] Note, it is possible that there was no change in the number of conflicts but there were less recusals. If this is correct, then inferences may be incorrect. Now I present the main theorem for this section.

Theorem 4.2. *Suppose assumptions 4.1, 4.3 and 4.4 hold. If* $X^\star + Y^\star = 8$ *then* $X|R = 1 \sim X$.

Thus I can measure the Kullback-Leibler divergence of $X|R = 0$ from X using the estimated Kullback-Leibler divergence of $X|R = 0$ from $X|R = 1$. Since $X|R = 1$ is unobserved, I replace it with its imputed version (from the previous section). The estimated Kullback-Leibler divergence of $X|R = 0$ from $X^{imputed}|R = 1$ is 0.049.

Now that I can observe the proxy distribution for X, I can generate W using structural parameters, ϕ_1 and ϕ_2, such that $KL(X; W) \approx KL(X; X|R = 0) = KL(X^{imputed}|R = 1; X|R = 0)$. Note that $KL(X; W) = KL(X; X|R = 0)$ does not necessarily imply $W \sim X|R = 0$. This feature is desirable since I am modeling only one potential mechanism and there are additional possible reasons why $X \not\sim X|R = 0$. One last assumption needs to be made before simulation. In this model, if a justice were to fail to recuse himself for all

Supreme Court Justices are appointed for life and no amount of negative media attention would hurt his tenure at the court (unless it was a controversy that could lead to impeachment).

[29] From 1988 to 2014 there were only 8 cases with more than 1 recusal.

possible X, it would look like he is randomly recusing himself and not being 'strategic' about his recusals.

Assumption 4.5. *If* $R = 0$ *and* $C_X + C_Y = 1$ *then* $X \in \{2, 3, 4, 5, 6, 7\}$

This assumption can be thought of as a foresight assumption. It says that a justice may fail to recuse himself properly if there might be some contention in the court (and hence his vote might make a difference in the outcome). If the court is going to be fairly unanimous in its decision ($X \in \{0, 1, 8, 9\}$), then the justice does not have much incentive to fail to recuse himself since it would not make a difference. Since recusal is typically done before arguments, the justice may have some idea how the others may vote, but does not have perfect foresight. Hence, why the values of 2,3,4,5,6 and 7 were chosen instead of just 4 and 5.

Define $\phi_1 \equiv Pr(C_X + C_Y = 1)$ and $\phi_2 \equiv Pr(R = 1|C_X + C_Y = 1)$ (note under Assumption 4.4 , $Pr(C_X + C_Y = 0) = 1 - \phi_1$). Using Assumptions 4.1,4.3,4.4 and 4.5, the simulation of W is as follows.

Simulation of W

1. Draw w' from $W' \sim multinomial(1, \eta)$
2. Draw from $C_X + C_Y$
3. If $C_X + C_Y = 0$ then store w' as a draw from W
4. If $C_X + C_Y = 1$ then draw from $R|(C_X + C_Y = 1)$
   (a) If $R|(C_X + C_Y = 1) = 0$ then store w' as a draw from W if $w' \in \{2, 3, 4, 5, 6, 7\}$
   (b) If $R|(C_X + C_Y = 1) = 1$ then discard w'

5. Repeat 1-4 10,000 times

The parameter η needs to be calibrated such that $Pr(X = i) = \eta_i$ for $i \in \{0, 1, ..., 9\}$. By Theorem 4.2, η can be calibrated using the maximum likelihood estimate from $X^{imputed}|R = 1$, which is $\eta = (.255\ .101\ .101\ .098\ .069\ .050\ .058\ .065\ .061\ .143)$. To find the appropriate ϕ_1 and ϕ_2, I run the above simulation using pairs of (ϕ_1, ϕ_2) over the grid $(\phi_1, \phi_2) \in [0, 1] \times [0, 1]$. I keep the pairs of (ϕ_1, ϕ_2) that generate the appropriate Kullback-Leibler divergence of 0.049.

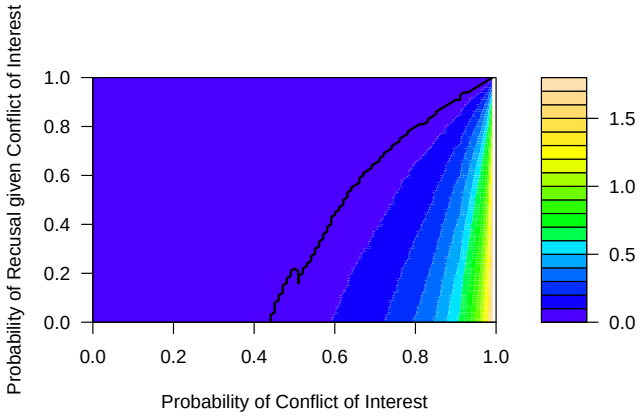

Figure 4.7: Estimated Kullback-Leibler divergences

The results of the simulation can best be shown in two figures. The first, Figure 4.7, is a contour plot showing the estimated Kullback- Leibler divergence of W from $X|R = 1$ for a given (ϕ_1, ϕ_2) pair. The x-axis is ϕ_1, the probability of there being a conflict of interest in a case. The y-axis is ϕ_2, the probability of recusal given there is a conflict of interest in a case. The colors represent the estimated Kullback-Leibler divergences of W from $X|R = 1$ for the given parameters. Darker colors are a smaller divergence and brighter colors are a larger divergence. The black line is a Kullback-Leibler divergence of $X|R = 0$ from $X|R = 1$ which is 0.051. I keep all pairs of (ϕ_1, ϕ_2) on the black line as potential parameters and discard all others.

The second, Figure 4.8, shows predictions arising from the selected potential parameters for a court that hears 100 cases in a year. The Supreme Court hears less than 100 cases a year, but 100 was chosen so that estimates can easily be interpreted in terms of percentage of cases. The x-axis is the number of cases that have a conflict of interest. The left y-axis is the

number of cases that have a conflict and a recusal and corresponds to the black line. The right y-axis is the number of cases that have a conflict but no recusal, it corresponds to the dashed red line. The two y-axes have the same scale. The pairs of (ϕ_1, ϕ_2) on the black line from Figure 4.7 trace out the black and dashed red lines on Figure 4.8. The vertical dotted black line represents where recusals equals 10. Every pair of parameters that predicts more than 10 recusals is not supported by the data because the court never has had more than 10% of cases with a recusal in a given term (see Figure 4.4). Thus only the values to the left of the dotted line are supported by the data. Table 4.4 shows these values.

Table 4.4 shows predictions from the simulation for a court that hears 100 cases a term. The first two columns are the parameters selected from the simulation that are supported by the data. The next three columns are the number of cases where a justice has a conflict of interest, the number of cases with a recusal, and the number of cases where a justice has a conflict of interest but does not recuse. The number of cases with a conflict of interest ranges from 45 to 57 and the number of recusals increases from 0 to 11, increasing with the number of conflicts. The number of cases with a conflict but no recusal hovers between 44 and 47. These are quite large estimates for the number of conflicted cases and cases with a conflict but no recusal. The reason is likely because I am only considering one mechanism for the observed data (strategic recusals) and I am ignoring all other possible selection mechanisms that could result in the observed data. This mechanism must then absorb all the other possible sources of selection. Additionally, this procedure results in a set of potential point estimates, it does not provide estimates of error. Therefore these results can be thought of as a rough estimate of the number of conflicts and conflicts without

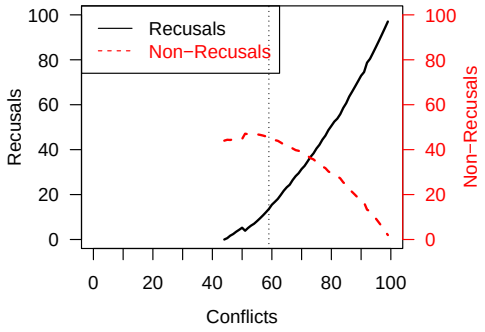

Figure 4.8: Simulated conflicts of interest and recusals

recusals.

ϕ_1	ϕ_2	#Conf	#Rec	#NonRec
.45	.01	45	0.6	44.4
.46	.04	46	1.7	44.3
.47	.05	47	2.4	44.6
.48	.07	48	3.5	44.5
.49	.09	49	4.3	44.7
.50	.10	50	5.2	44.8
.51	.08	51	3.9	47.1
.52	.10	52	5.1	46.9
.53	.12	53	6.2	46.8
.54	.13	54	7.0	47.0
.55	.15	55	8.2	46.8
.56	.17	56	9.4	46.6
.57	.19	57	10.7	46.3

Table 4.4: Simulation results

I perform an alternative specification where the support for X is constrained by combining $0, 1$ and $3, 4, 5, 6$ and $8, 9$ and the counts are constrained to terms after 1988 in accordance with the robustness checks in the previous section. This specification is also more favorable to Assumption 4.4 because there were less recusals after 1988 which could be explained by there being a low number of conflicts. By restricting the support and the counts, the estimated divergence is 0.044 and the calibrated η is $\eta = (.250\ .080\ .074\ .069\ .074\ .064\ .064\ .074\ .037\ .213)$. The simulation is ran the same as before except there is an additional step between 5 and 6 where the values from W are combined according to the support restriction. The results are presented in Table 4.5. This specification resulted in a slightly more conservative estimate. The number of cases where a justice has a conflict ranges from 47 to 56 cases. The number of cases where a justice has a conflict but does not recuse himself ranges from 45 to 46 cases. Again, these results should be interpreted as a rough estimate.

ϕ_1	ϕ_2	#Conf	#Rec	#NonRec
.47	.00	47	0.0	47
.48	.01	48	0.7	47.3
.49	.03	49	1.4	47.6
.50	.05	50	2.5	47.5
.51	.09	51	4.4	46.6
.52	.10	52	5.4	46.6
.53	.13	53	6.7	46.3
.54	.15	54	7.9	46.1
.55	.16	55	9.1	45.9
.56	.18	56	1.4	45.7

Table 4.5: Simulation results with robustness check

4.4 Discussion

It is plausible that Supreme Court Justices might not always recuse themselves when they are supposed to. Some studies in the past did provide evidence that justices recuse themselves strategically. Compared to previous literature this paper uses a structural approach to investigate strategic recusals and agrees there is evidence that Supreme Court Justices sometimes recuse themselves strategically. Under certain assumptions, this paper finds that the percent of cases with conflict of interest ranges from 45% to 57%, it follows that about 44% to 47% of cases will have a conflict of interest but no recusal. Future research could bring more precision to these estimates. Additionally, there is likely to be some changes in these results by further exploiting heterogeneity in justices, types of cases, and time. Lastly, future research could explore the question of "is strategic recusal bad?" and "how bad is it?"

Bibliography

Akaike, H. (1998): *Information Theory and an Extension of the Maximum Likelihood Principle*, New York, NY: Springer New York, 199–213.

Alhamzawi, R., K. Yu, and D. F. Benoit (2012): "Bayesian adaptive lasso quantile regression," *Statistical Modelling*, 12, 279–297.

Andrews, D. W. K. (1991): "Heteroskedasticity and autocorrelation consistent covariance matrix estimation," *Econometrica*, 59, 817–858.

Arrington, T. S. and S. Brenner (2004): "Strategic voting for damage control on the supreme court," *Political Research Quarterly*, 57, 565–573.

Benoit, D. F., R. Alhamzawi, K. Yu, and D. Van den Poel (2014): *BayesQR: Bayesian quantile regression*, R package version 2.2.

Benoit, D. F. and D. Van den Poel (2012): "Binary quantile regression: a Bayesian approach based on the asymmetric Laplace distribution," *Journal of Applied Econometrics*, 27, 1174–1188.

Bera, A. K. (1984): "The use of linear approximation to nonlinear regression analysis," *Sankhy: The Indian Journal of Statistics, Series B (1960-2002)*, 46, 285–290.

Berk, R. H. (1966a): "Correction notes: correction to limiting behavior of posterior distributions when the model is incorrect," *The Annals of Mathematical Statistics*, 37, 745–746.

——— (1966b): "Limiting behavior of posterior distributions when the model is incorrect," *The Annals of Mathematical Statistics*, 37, 51–58.

——— (1970): "Consistency a posteriori," *The Annals of Mathematical Statistics*, 41, 894–906.

Bhat, C. R. (1995): "A heteroscedastic extreme value model of intercity travel mode choice," *Transportation Research Part B: Methodological*, 29, 471 – 483.

Black, R. and L. Epstein (2005): "Recusals and the problem of an equally divided supreme court," *Journal of Applied Practices & Process*, 7, 75.

Buja, A., R. Berk, L. Brown, E. George, A. K. Kuchibhotla, L. Zhao, and K. Zhang (2016a): "Models as approximations part II: a general theory of model-robust regression," Unpublished.

Buja, A., R. Berk, L. Brown, E. George, E. Pitkin, M. Traskin, L. Zhao, and K. Zhang (2016b): "Models as approximations a conspiracy of random regressors and model deviations against classical inference in regression," Unpublished.

Bunke, O. and X. Milhaud (1998): "Asymptotic behavior of Bayes estimates under possibly incorrect models," *The Annals of Statistics*, 26, 617–644.

Chernozhukov, V. and H. Hong (2003): "An MCMC approach to classical estimation," *Journal of Econometrics*, 115, 293 – 346.

Choi, H. and N. M. Kiefer (2011): "Geometry of the log-likelihood ratio statistic in misspecified models," *Journal of Statistical Planning and Inference*, 141, 2091 – 2099.

Chow, G. C. (1984): "Maximum-likelihood estimation of misspecified models," *Economic Modelling*, 1, 134 – 138.

Cramer, J. (2005): "Omitted variables and misspecified disturbances in the logit model," Tinbergen Institute Discussion Papers 05-084/4, Tinbergen Institute.

Dagpunar, J. (1989): "An easily implemented generalised inverse Gaussian generator," *Communications in Statistics - Simulation and Computation*, 18, 703–710.

Davidson, R. and J. G. MacKinnon (1984): "Convenient specification tests for logit and probit models," *Journal of Econometrics*, 25, 241 – 262.

De Blasi, P. and S. G. Walker (2013): "Bayesian asymptotics with misspecified models," *Statistica Sinica*, 169–187.

Deegan Jr., J. (1976): "The consequences of model misspecification in regression analysis," *Multivariate Behavioral Research*, 11, 237–248.

Diaconis, P. and D. A. Freedman (1986a): "On inconsistent Bayes estimates of location," *The Annals of Statistics*, 14, 68–87.

——— (1986b): "On the consistency of Bayes estimates," *The Annals of Statistics*, 14, 1–26.

Domowitz, I. and H. White (1982): "Misspecified models with dependent observations," *Journal of Econometrics*, 20, 35 – 58.

Doob, J. L. (1949): "Application of the theory of martingales," *Le calcul des probabilites et ses applications*, 23–27.

Drovandi, C. C. and A. N. Pettitt (2011): "Likelihood-free Bayesian estimation of multivariate quantile distributions," *Computational Statistics & Data Analysis*, 55, 2541–2556.

Dubin, J. A. and L. Zeng (1991): "The heterogeneous logit model," *California Institute of Technology, Social Science Working Paper 759*.

Dumont, J. and J. Keller (2015): *RSGHB: functions for hierarchical Bayesian estimation: a flexible approach*, R package version 1.1.2.

Dutta, S., A. K. Ghosh, P. Chaudhuri, et al. (2011): "Some intriguing properties of Tukeys half-space depth," *Bernoulli*, 17, 1420–1434.

Eicker, F. (1967): "Limit theorems for regressions with unequal and dependent errors," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, Berkeley, Calif.: University of California Press, 59–82.

Eicker, F. et al. (1963): "Asymptotic normality and consistency of the least squares estimators for families of linear regressions," *The Annals of Mathematical Statistics*, 34, 447–456.

Embrechts, P. and M. Hofert (2013): "A note on generalized inverses," *Mathematical Methods of Operations Research*, 77, 423–432.

Epstein, L., W. Landes, and R. Posner (2013): *The behavior of federal judges: a theoretical and empirical study of rational choice*, Harvard University Press.

Feng, C., H. Wang, X. M. Tu, and J. Kowalski (2012): "A note on generalized inverses of distribution function and quantile transformation," *Applied Mathematics*.

Feng, Y., Y. Chen, and X. He (2015): "Bayesian quantile regression with approximate likelihood," *Bernoulli*, 21, 832–850.

Fernández-Villaverde, J. and J. F. Rubio-Ramírez (2004): "Comparing dynamic equilibrium models to data: a Bayesian approach," *Journal of Econometrics*, 123, 153 – 187.

Finn, J. D. and C. M. Achilles (1990): "Answers and questions about class size: a statewide experiment," *American Educational Research Journal*, 27, 557–577.

Folger, J. and C. Breda (1989): "Evidence from project STAR about class size and student achievement," *Peabody Journal of Education*, 67, 17–33.

Foutz, R. V. and R. C. Srivastava (1977): "The performance of the likelihood ratio test when the model is incorrect," *The Annals of Statistics*, 5, 1183–1194.

——— (1978): "The asymptotic distribution of the likelihood ratio when the model is incorrect," *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 6, 273–279.

Fox, M. and H. Rubin (1964): "Admissibility of quantile estimates of a single location parameter," *Annals of Mathematical Statistics*, 35, 1019–1030.

Freedman, D. A. (1963): "On the asymptotic behavior of Bayes estimates in the discrete case," *The Annals of Mathematical Statistics*, 34, 1386–1403.

——— (1965): "On the asymptotic behavior of Bayes estimates in the discrete case II," *The Annals of Mathematical Statistics*, 36, 454–456.

——— (2006): "On the so-called "Huber sandwich estimator" and "robust standard errors"," *The American Statistician*, 60, 299–302.

Freedman, D. A. and P. Diaconis (1983): "On inconsistent Bayes estimates in the discrete case," *The Annals of Statistics*, 11, 1109–1118.

Gartner, S. S. and G. M. Segura (2000): "Race, casualties, and opinion in the Vietnam war," *Journal of Politics*, 62, pp. 115–146.

Ghosal, S. (1997): "A review of consistency and convergence of posterior distribution," in *Varanashi Symposium in Bayesian Inference*, Banaras Hindu University.

Gould, E. D., V. Lavy, and M. D. Paserman (2004): "Immigrating to opportunity: estimating the effect of school quality using a natural experiment on ethiopians in israel," *The Quarterly Journal of Economics*, 119, 489–526.

Gourieroux, C., A. Monfort, and A. Trognon (1984a): "Pseudo maximum likelihood methods: applications to Poisson models," *Econometrica*, 52, 701–720.

——— (1984b): "Pseudo maximum likelihood methods: theory," *Econometrica*, 52, 681–700.

Griliches, Z. (1957): "Specification bias in estimates of production functions," *Journal of farm economics*, 39, 8–20.

Haan, P. and A. Uhlendorff (2006): "Estimation of multinomial logit models with unobserved heterogeneity using maximum simulated likelihood," *Stata Journal*, 6, 229–245(17).

Hallin, M., D. Paindaveine, and M. Šiman (2010): "Multivariate quantiles and multiple-output regression quantiles: from L1 optimization to halfspace depth," *Annals of Statistics*, 635–703.

Hausman, J. A. (1978): "Specification tests in econometrics," *Econometrica*, 46, 1251–1271.

Hinkley, D. V. (1969): "On the ratio of two correlated normal random variables," *Biometrika*, 56, 635–639.

——— (1970): "Correction: 'On the ratio of two correlated normal random variables'," *Biometrika*, 57, 683.

Hoff, P. and J. Wakefield (2013): "Bayesian sandwich posteriors for pseudo-true parameters," *Journal of Statistical Planning and Inference*, 143, 1638 – 1642.

Hole, A. R. (2006): "Small-sample properties of tests for heteroscedasticity in the conditional logit model," *Economics Bulletin*, 3, 1–14.

——— (2007): "Fitting mixed logit models by using maximum simulated likelihood," *Stata Journal*, 7, 388–401(14).

Hong, H. and B. Preston (2012): "Bayesian averaging, prediction and nonnested model selection," *Journal of Econometrics*, 167, 358 – 369, fourth Symposium on Econometric Theory and Applications (SETA).

Huber, P. (1967): "The behavior of the maximum likelihood estimates under nonstandard conditions," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley: University of California Press, vol. 1, 221–233.

Hume, R. J. (2014): "Deciding not to decide: the politics of recusals on the U.S. Supreme Court," *Law and Society Review*, 48, 621–655.

III, L. J. V. (2011): "The (un)constitutionality of supreme court recusal standards," *Wisconsin Law Review*, 2011.

Ingram, D. (2009): "Federal judges push back against recusal proposals, congress considers revising rules on judge disqualifications," *National Law Journal*.

Jacobs, D. and J. T. Carmichael (2002): "The political sociology of the death penalty: a pooled time-Series analysis," *American Sociological Review*, 67, pp. 109–131.

Jeliazkov, I. and E. H. Lee (2010): "MCMC perspectives on simulated likelihood estimation," *Advances in Econometrics*, 26, 3,39.

Johnson, T. R., J. F. Spriggs, and P. J. Wahlbeck (2005): "Passing and strategic voting on the U.S. Supreme Court," *Law and Society Review*, 39, 349–378.

Keane, M. (1992): "A note on identification in the multinomial probit model," *Journal of Business and Economic Statistics*, 10, 193–200.

Khare, K. and J. P. Hobert (2012): "Geometric ergodicity of the Gibbs sampler for Bayesian quantile regression," *Journal of Multivariate Analysis*, 112, 108 – 116.

Kleijn, B. and A. Van der Vaart (2012): "The Bernstein-von-Mises theorem under misspecification," *Electronic Journal of Statistics*, 6, 354–381.

Kleijn, B. J. and A. van der Vaart (2006): "Misspecification in infinite-dimensional Bayesian statistics," *The Annals of Statistics*, 837–877.

Koenker, R. (2005): *Quantile regression*, 38, Cambridge university press.

Koenker, R. and G. Bassett (1978): "Regression quantiles," *Econometrica*, 33–50.

Kong, L. and I. Mizera (2012): "Quantile tomography: using quantiles with multivariate data," *Statistica Sinica*, 22, 1589–1610.

Koop, G. and D. J. Poirier (1993): "Bayesian analysis of logit models using natural conjugate priors," *Journal of Econometrics*, 56, 323 – 340.

Kottas, A. and M. Krnjajić (2009): "Bayesian semiparametric modelling in quantile regression," *Scandinavian Journal of Statistics*, 36, 297–319.

Kotz, S., T. Kozubowski, and K. Podgorski (2001): *The Laplace distribution and generalizations: a revisit with applications to communications, economics, engineering, and finance*, Progress in Mathematics Series, Birkhäuser Boston.

Kozumi, H. and G. Kobayashi (2011): "Gibbs sampling methods for Bayesian quantile regression," *Journal of Statistical Computation and Simulation*, 81, 1565–1578.

Krueger, A. B. (1999): "Experimental estimates of education production functions," *The Quarterly Journal of Economics*, 114, 497–532.

Kullback, S. and R. A. Leibler (1951): "On information and sufficiency," *Annals of Mathematical Statistics*, 22, 49–86.

Laine, B. (2001): "Depth contours as multivariate quantiles: a directional approach," Master's thesis, Univ. Libre de Bruxelles, Brussels.

Lancaster, T. and S. Jae Jun (2010): "Bayesian quantile regression methods," *Journal of Applied Econometrics*, 25, 287–307.

Lange, K. (1999): *Numerical Analysis for Statisticians*, Statistics and computing, Springer.

Lassen, D. D. (2005): "The effect of information on voter turnout: evidence from a natural experiment," *American Journal of Political Science*, 49, pp. 103–118.

Lee, J. and S. N. MacEachern (2011): "Consistency of Bayes estimators without the assumption that the model is correct," *Journal of Statistical Planning and Inference*, 141, 748 – 757.

Long, J. S. and L. H. Ervin (2000): "Using heteroscedasticity consistent standard errors in the linear regression model," *The American Statistician*, 54, 217–224.

Luce, R. D. (1959): *Individual choice behavior: a theoretical analysis*, Wiley.

Lv, J. and J. S. Liu (2014): "Model selection principles in misspecified models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76, 141–167.

MacKinnon, J. G. and H. White (1985): "Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties," *Journal of Econometrics*, 29, 305 – 325.

Marsaglia, G. (1965): "Ratios of normal variables and ratios of sums of uniform variables," *Journal of the American Statistical Association*, 60, 193–204.

——— (2006): "Ratios of normal variables," *Journal of Statistical Software*, 16.

Mcfadden, D. (1974): "Conditional logit analysis of qualitative choice behavior," in *Frontiers in econometrics*, ed. by P. Zarembka, New York: Academic Press, 105–142.

McFadden, D. and K. Train (2000): "Mixed MNL models for discrete response," *Journal of Applied Econometrics*, 15, 447–470.

Monfort, A. (1996): "A reappraisal of misspecified econometric models," *Econometric Theory*, 12, 597–619.

Mosteller, F. (1995): "The Tennessee study of class size in the early school grades," *Future of Children*, 5, 113–127.

Müller, U. K. (2013): "Risk of Bayesian inference in misspecified models, and the sandwish covariance matrix," *Econometrica*, 81, 1805–1849.

Myers, R. H. and S. J. Lahoda (1975): "A generalization of the response surface mean square error criterion with a specific application to the scope," *Technometrics*, 17, 481–486.

Newey, W. K. (1987): "Generic uniqueness of population quasi maximum likelihood parameters," Unpublished.

Newey, W. K. and D. G. Steigerwald (1997): "Asymptotic bias for quasi-maximum-likelihood estimators in conditional heteroskedasticity models," *Econometrica*, 65, 587–599.

Newey, W. K. and K. D. West (1987): "A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix," *Econometrica*, 55, 703–708.

Paindaveine, D. and M. Šiman (2011): "On directional multiple-output quantile regression," *Journal of Multivariate Analysis*, 102, 193 – 212.

Rahman, M. A. (2016): "Bayesian quantile regression for ordinal models," *Bayesian Analysis*, 11, 1–24.

Ramalho, E. A. and J. J. Ramalho (2010): "Is neglected heterogeneity really an issue in binary and fractional regression models? A simulation exercise for logit, probit and loglog models," *Computational Statistics and Data Analysis*, 54, 987 – 1001.

Ramamoorthi, R., K. Sriram, R. Martin, et al. (2015): "On posterior concentration in misspecified models," *Bayesian Analysis*, 10, 759–789.

Ramsey, J. B. (1969): "Tests for specification errors in classical linear least-squares regression analysis," *Journal of the Royal Statistical Society. Series B (Methodological)*, 31, 350–371.

Ramsey, J. B. and P. Schmidt (1976): "Some further results on the use of OLS and BLUS residuals in specification error tests," *Journal of the American Statistical Association*, 71, 389–390.

Rao, P. (1971): "Some notes on misspecification in multiple regressions," *The American Statistician*, 25, 37–39.

Rice, J. K. (2010): "The impact of teacher experience: examining the evidence and policy implications. Brief no. 11." *National Center for Analysis of Longitudinal Data in Education Research*.

Robert, C. (1991): "Generalized inverse normal distributions," *Statistics & Probability Letters*, 11, 37–41.

Rousseeuw, P. J. and I. Ruts (1999): "The depth function of a population distribution," *Metrika*, 49, 213–244.

Ruud, P. A. (1983): "Sufficient conditions for the consistency of maximum likelihood estimation despite misspecification of distribution in multinomial discrete choice models," *Econometrica*, 51, 225–228.

——— (1996): "Simulation of the multinomial probit model: an analysis of covariance matrix estimation," Working paper.

Sample, J. J. (2013): "Supreme court recusal: from Marbury to the modern day," *Georgetown Journal of Legal Ethics*, 26, 95.

Schwartz, L. (1965): "On Bayes procedures," *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 4, 10–26.

Serfling, R. (2002): "Quantile functions for multivariate analysis: approaches and applications," *Statistica Neerlandica*, 56, 214–232.

Serfling, R. and Y. Zuo (2010): "Discussion," *Annals of Statistics*, 38, 676–684.

Shalizi, C. R. et al. (2009): "Dynamics of Bayesian updating with dependent data and misspecified models," *Electronic Journal of Statistics*, 3, 1039–1074.

Small, C. G. (1990): "A survey of multidimensional medians," *International Statistical Review / Revue Internationale de Statistique*, 58, 263–277.

Spriggs, J. F. I., F. Maltzman, and P. J. Wahlbeck (1999): "Bargaining on the U.S. Supreme Court: justices' responses to majority opinion drafts," *Journal of Politics*, 61, 485–506.

Sriram, K., R. Ramamoorthi, and P. Ghosh (2013): "Posterior consistency of Bayesian quantile regression based on the misspecified asymmetric Laplace density," *Bayesian Analysis*, 8, 479–504.

Stempel, J. W. (2009): "Chief William's ghost: the problematic persistence of the duty to sit doctrine," *Scholarly Works*.

Taddy, M. A. and A. Kottas (2010): "A Bayesian nonparametric approach to inference for quantile regression," *Journal of Business & Economic Statistics*, 28.

Theil, H. (1957): "Specification errors and the estimation of economic relationships," *Revue de l'Institut International de Statistique*, 41–51.

Thompson, P., Y. Cai, R. Moyeed, D. Reeve, and J. Stander (2010): "Bayesian nonparametric quantile regression using splines," *Computational Statistics & Data Analysis*, 54, 1138–1150.

Train, K. E. (2009): *Discrete Choice Methods with Simulation*, Cambridge Books, Cambridge University Press.

Tukey, J. W. (1975): "Mathematics and the picturing of data," in *Proceedings of the 1975 International Congress of Mathematics*, vol. 2, 523–531.

Virelli, L. J. (2012): "Congress, the constitution, and supreme court recusal," *Washington and Lee Law Review*, 69.

Wahlbeck, P. J. (2006): "Strategy and constraints on supreme court opinion assignment," *University of Pennsylvania Law Review*, 154, pp. 1729–1755.

Waldmann, E. and T. Kneib (2015): "Bayesian bivariate quantile regression," *Statistical Modelling*.

Walker, S. G. (2013): "Bayesian inference with misspecified models," *Journal of Statistical Planning and Inference*, 143, 1621–1633.

Wang, Y. H. (1993): "On the number of successes in independent trials," *Statistica Sinica*, 3, 295–312.

Watson, J., C. Holmes, et al. (2016): "Approximate models and robust decisions," *Statistical Science*, 31, 465–489.

Wheeler, R. R. (2014): "A primer on regulating federal judicial ethics," *Arizona Law Review*, 53.

White, H. (1980a): "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity," *Econometrica: Journal of the Econometric Society*, 817–838.

——— (1980b): "Using least squares to approximate unknown regression functions," *International Economic Review*, 149–170.

——— (1981): "Consequences and detection of misspecified nonlinear regression models," *Journal of the American Statistical Association*, 76, 419–433.

——— (1982): "Maximum likelihood estimation of misspecified models," *Econometrica: Journal of the Econometric Society*, 1–25.

——— (1983): "Corrigendum [maximum likelihood estimation of misspecified models]," *Econometrica*, 51, 513.

Word, E., J. Johnston, H. P. Bain, B. D. Fulton, J. B. Zaharias, C. M. Achilles, M. N. Lintz, J. Folger, and C. Breda (1990): "The state of Tennessee's Student/Teacher Achievement Ratio (STAR) project: technical report 1985 – 1990," Tech. rep., Tennessee State Department of Education.

Yang, Y., H. J. Wang, and X. He (2016): "Posterior inference in Bayesian quantile regression with asymmetric Laplace likelihood," *International Statistical Review*, 84, 327–344.

Yatchew, A. and Z. Griliches (1985): "Specification error in probit models," *Review of Economics and Statistics*, 67, 134–39.

Yu, K., Z. Lu, and J. Stander (2003): "Quantile regression: applications and current research areas," *Journal of the Royal Statistical Society: Series D (The Statistician)*, 52, 331–350.

Yu, K. and R. A. Moyeed (2001): "Bayesian quantile regression," *Statistics & Probability Letters*, 54, 437 – 447.

Yu, K. and J. Zhang (2005): "A three-parameter asymmetric Laplace distribution and its Extension," *Communications in Statistics - Theory and Methods*, 34, 1867–1879.

Appendix A

Chapter 1

A.1 Standard error and MSE under omitted stochastic variables

Rao (1971) investigates the properties of ordinary least squares estimators when the truth is $Y = \alpha_1 X_1 + ... + \alpha_k X_k + \alpha_{k+1} X_{k+1} + \epsilon$ but a misspecified model $Y = \beta_1 X_1 + ... + \beta_k X_k + \eta$ is used. He derives results for the effects on the standard errors and mean square error. However he assumes regressors are fixed which is usually not the case for economic models. In this appendix I update his results for when regressors are stochastic. Define $\sigma^2_\epsilon = Var(Y|X_1, ..., X_k, X_{k+1})$ and $\sigma^2_\eta = Var(Y|X_1, ..., X_k)$. Then by the law of total variance $\sigma^2_\eta = E(Var(Y|X_1, ..., X_k, X_{k+1})|X_1, ..., X_k) + Var(E(Y|X_1, ..., X_k, X_{k+1})|X_1, ..., X_k) = \sigma^2_\epsilon + \alpha^2_{k+1} Var(X_{k+1}|X_1, ..., X_k)$. Define $S^2_{1.23...k\cdot}$ as the residual sum of squares auxiliary regression with X_1 as the dependent variable and $(X_2, ..., X_k, X_{k+1})$ as the independent variables. Then

$$Var(\hat{\beta}) = (\sigma^2_\epsilon + \alpha^2_{k+1} Var(X_{k+1}|X_1, ..., X_k))(X'X)^{-1}$$

and

$$Var(\hat{\beta}_1) = (\sigma_\epsilon^2 + \alpha_{k+1}^2 Var(X_{k+1}|X_1, ..., X_k))/S_{1.23...k}^2.$$

Thus $Var(\hat{\beta}_1) \leq Var(\hat{\alpha}_1)$ is true when $\alpha_{k+1}^2 Var(X_{k+1}|X_1, ..., X_k) \leq \sigma_\epsilon^2 r_{1,k+1.23...k}^2/(1 - r_{1,k+1.23...k}^2)$, where $r_{1,k+1.23...k}^2$ is the partial correlation between X_1 and X_{k+1} keeping $(X_2, ..., X_k)$ constant. Using these results we the effect on standard errors when variables are omitted.

Theorem A.1. *In the classical linear regression model, omission of a variable specified by the truth decreases the variance of the least squares estimator for the coefficient of the first covariate provided sufficiently large σ_ϵ^2 and $r_{1,k+1.23...k}^2$ and sufficiently small α_{k+1}^2 and $Var(X_{k+1}|X_1, ..., X_k)$.*

Mimicking the derivation in Rao (1971) we get

$$MSE(\hat{\beta}_1) = \alpha_{k+1}^2 \cdot b_{k+1,1.23...k}^2 + (\sigma_\epsilon^2 + \alpha_{k+1}^2 Var(X_{k+1}|X_1, ..., X_k))/S_{1.23...k}^2,$$

$$\alpha_{k+1}^2 \leq \frac{\sigma_\epsilon^2}{S_{k+1,123...k}^2 + \frac{1-r_{1,k+1.23...k}^2}{r_{1,k+1.23...k}^2} Var(X_{k+1}|X_1, ..., X_k)}$$

and

$$|\alpha_{k+1}| \leq (Var(\hat{\alpha}_{k+1})^{-1} + \frac{1 - r_{1,k+1.23...k}^2}{\sigma_\epsilon^2 r_{1,k+1.23...k}^2} Var(X_{k+1}|X_1, ..., X_k))^{-1/2}.$$

Using these results we find the effect on MSEs when variables are omitted.

Theorem A.2. *In the classical linear regression model, discarding an independent vari-*

able decreases the mean square error of the least squares estimator for the coefficient of the first covariate provided sufficiently large σ_ϵ^2, $Var(\hat{\alpha}_{k+1})$ *and* $r^2_{1,k+1.23...k}$ *and sufficiently small* $|\alpha_{k+1}|$ *and* $Var(X_{k+1}|X_1, ..., X_k)$.

Appendix B

Chapter 2

B.1 Lemmas and proofs

The proof of Lemma 2.1 is below.

Proof. The strategy of this proof is to first separate the log and focus on one half of it (the denominator). Then evaluate the first $N-1$ summations at their first term and fully evaluate the last summation. With this last one fully evaluated, I show that a recursive pattern appears that can be used to find the desired result.

For ease of readability I replace P_{nj} with p_{nj} and replace P^0_{nj} with q_{nj}. It is clear the summations can be reordered and the KL divergence can be written as

$$KL(G||F) = \sum_{y_N \in Y_N} \cdots \sum_{y_1 \in Y_1} log\left[\prod_{n=1}^{N}\prod_{j=1}^{J} q_{nj}^{y_{nj}}\right] \prod_{n=1}^{N}\prod_{j=1}^{J} q_{nj}^{y_{nj}} - log\left[\prod_{n=1}^{N}\prod_{j=1}^{J} p_{nj}^{y_{nj}}\right] \prod_{n=1}^{N}\prod_{j=1}^{J} q_{nj}^{y_{nj}}.$$

I will focus on

$$C_N \equiv -\sum_{y_N \in Y_N} \cdots \sum_{y_1 \in Y_1} log\left[\prod_{n=1}^{N}\prod_{j=1}^{J} p_{nj}^{y_{nj}}\right]\prod_{n=1}^{N}\prod_{j=1}^{J} q_{nj}^{y_{nj}}.$$

The result for the other half with follows by analogy. Define $C_N^\star = \sum_{n=1}^{N}\sum_{j=1}^{J} log(p_{nj})q_{nj}$. I wish to show $-C_N = C_N^\star$. Without loss of generality, I focus on the first term of each of the sums Y_N, Y_{N-1},..., Y_2. This is equivalent to focusing on $y_i = [1, 0, ..., 0]$ for $i \in \{N, N-1, ..., 2\}$ and putting all other terms in the sums into a variable called $C^{(1)}$. Then focusing on evaluating the summation over Y_1, $-C_N$ becomes

$$\begin{aligned}
-C_N &= log(p_{N1}p_{(N-1)1}...p_{21}p_{11})q_{N1}q_{(N-1)1}...q_{21}q_{11} \\
&\quad + log(p_{N1}p_{(N-1)1}...p_{21}p_{12})q_{N1}q_{(N-1)1}...q_{21}q_{12} \\
&\quad + ... \\
&\quad + log(p_{N1}p_{(N-1)1}...p_{21}p_{12})q_{N1}q_{(N-1)1}...q_{21}q_{1J} + C^{(1)}
\end{aligned}$$

where $C^{(1)}$ is the rest of the terms in the summations for $Y_N, Y_{N-1}, ..., Y_1$. To be explicit $C^{(1)} = \sum_{y_N \in Y_N'} \cdots \sum_{y_2 \in Y_2'}\sum_{y_1 \in Y_1'} log\left[\prod_{n=1}^{N}\prod_{j=1}^{J} p_{nj}^{y_{nj}}\right]\prod_{n=1}^{N}\prod_{j=1}^{J} q_{nj}^{y_{nj}}$ where
$Y_N' \times ... \times Y_2' \times Y_1' = Y_N \times ... \times Y_2 \times Y_1 / \{[1, 0, ..., 0] \times ... \times [1, 0, ..., 0] \times Y_1\}$. $-C_N$ can be re-written as

$$\begin{aligned}
-C_N &= log(p_{N1}p_{(N-1)1}...p_{21})q_{N1}q_{(N-1)1}...q_{21}(q_{11} + q_{12} + ... + q_{1J}) \\
&\quad + q_{N1}q_{(N-1)1}...q_{21}(q_{11}log(p_{11}) + q_{12}log(p_{12}) + ... + q_{1J}log(p_{1J})) + C^{(1)} \\
&= log(p_{N1}p_{(N-1)1}...p_{21})q_{N1}q_{(N-1)1}...q_{21} + q_{N1}q_{(N-1)1}...q_{21}C_1^\star + C^{(1)}
\end{aligned}$$

The recursive pattern now exists. I can make this more clear by evaluating the next summation for Y_2. With the summation for Y_1 evaluated, the summation over Y_2 focusing on

$y_i = [1, 0, ..., 0]$ for $i \in \{N, N-1, ..., 3\}$ becomes

$$\begin{aligned}
-C_N &= log(p_{N1}p_{(N-1)1}...p_{21})q_{N1}p_{(N-1)1}...p_{31}(q_{21} + q_{22} + ... + q_{2J}) \\
&\quad + q_{N1}q_{(N-1)1}...q_{31}(C_1^\star + q_{21}log(p_{21}) + q_{22}log(p_{22}) + ... + q_{2J}log(p_{2J})) + C^{(2)} \\
&= log(p_{N1}p_{(N-1)1}...p_{31})q_{N1}p_{(N-1)1}...p_{31} + q_{N1}q_{(N-1)1}...q_{31}C_2^\star + C^{(2)}
\end{aligned}$$

Where $C^{(2)} = \sum_{y_N \in Y_N'} \cdots \sum_{y_2 \in Y_2'} \sum_{y_1 \in Y_1'} log \left[\prod_{n=1}^{N} \prod_{j=1}^{J} p_{nj}^{y_{nj}} \right] \prod_{n=1}^{N} \prod_{j=1}^{J} q_{nj}^{y_{nj}}$ where
$Y_N' \times ... \times Y_2' \times Y_1' = Y_N \times ... \times Y_2 \times Y_1 / \{[1, 0, ..., 0] \times ... \times Y_2 \times Y_1\}$. Notice that $C^{(N)} = 0$ since it is a summation over nothing. By iterating on this for $Y_1, Y_2, ..., Y_{(N-1)}$ the last summation is

$$\begin{aligned}
-C_N &= log(p_{N1})q_{N1}(q_{(N-1)1} + q_{(N-1)2} + ... + q_{(N-1)J}) \\
&\quad + q_{N1}(C_{N-2}^\star + q_{(N-1)1}log(p_{(N-1)1}) + q_{(N-1)2}log(p_{(N-1)2}) + ... + q_{(N-1)J}log(p_{(N-1)J})) \\
&\quad + C^{(N-1)} \\
&= log(p_{N1})q_{N1} + q_{N1}C_{N-1}^\star + C^{(N-1)} \\
&= log(p_{N1})q_{N1} + q_{N1}C_{N-1}^\star + log(p_{N2})q_{N2} + q_{N2}C_{N-1}^\star + ... + log(p_{NJ})q_{NJ} + q_{NJ}C_{N-1}^\star \\
&= log(p_{N1})q_{11} + log(p_{N2})q_{N2} + ... + log(p_{NJ})q_{NJ} + C_{N-1}^\star(q_{N1} + q_{N2} + ... + q_{NJ}) \\
&= log(p_{N1})q_{N1} + log(p_{N2})q_{N2} + ... + log(p_{NJ})q_{NJ} + C_{N-1}^\star \\
&= C_N^\star
\end{aligned}$$

By symmetry a similar result holds for the numerator of the log and the desired result follows. □

The proof of Theorem 2.1 is below.

Proof. Note that $KL(G||F)$ is differentiable with respect to the β vector. Since the choice

probability $P_{nj}(\beta)$ is log-concave (see Mcfadden (1974)), a scaler multiplied to a concave function is concave and a sum of concave functions is concave, then the β vector that minimizes the KL divergence is a unique minimum. Thus $\beta^\star$ is the solution to $W(\beta)|_{\beta=\beta^\star} \equiv \frac{d}{d\beta}KL(G||F)\Big|_{\beta=\beta^\star} = 0$, call this $W(\beta)$. Thus the unique minimizer is the solution to the following equations

$$
\begin{aligned}
W(\beta) &= \frac{d}{d\beta}KL(G||F) \\
&= \sum_{n=1}^{N}\sum_{j=1}^{J} P_{nj}(\beta)P_{nj}^0 \begin{bmatrix} \sum_{i=1}^{J}(X_{ni}-X_{nj})^{(1)}e^{(X_{ni}-X_{nj})\beta} \\ \vdots \\ \sum_{i=1}^{J}(X_{ni}-X_{nj})^{(p)}e^{(X_{ni}-X_{nj})\beta} \end{bmatrix} \\
&= \sum_{n=1}^{N}\sum_{j=1}^{J} P_{nj}^0 \begin{bmatrix} \left[\sum_{i=1}^{J} X_{ni}^{(1)}P_{ni}(\beta)\right] - X_{nj}^{(1)} \\ \vdots \\ \left[\sum_{i=1}^{J} X_{ni}^{(p)}P_{ni}(\beta)\right] - X_{nj}^{(p)} \end{bmatrix} \\
&= \sum_{n=1}^{N}\sum_{j=1}^{J} (P_{nj}(\beta) - P_{nj}^0) \begin{bmatrix} X_{nj}^{(1)} \\ \vdots \\ X_{nj}^{(p)} \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}
\end{aligned}
$$

where $p \equiv length(\beta)$ and $X_{ni}^{(k)}$ and $(X_{ni}-X_{nj})^{(k)}$ is the kth vector of the matrix X_{ni} and $(X_{ni}-X_{nj})$ respectively. [1] □

Using Assumption 2.3, I show the mixed logit can approximate any random utility model.

Lemma B.1. *Suppose Assumption 2.3 holds, then the resulting choice probability can be approximated to any degree of accuracy by* $P_{nj} \approx \int \frac{e^{X_{nj}\beta+Z_{nj}b_n}}{\sum_{i=1}^{J} e^{X_{ni}\beta+Z_{ni}b_n}} dF_b$ *for some* Z_{nj} *and* $b_n \sim F_b$.

[1] Notice that since the summation does not necessarily converge, the minimizer does not might not have any stable asymptotic behavior. However, I conjecture that under some mild conditions it can be shown to converge.

Proof. This proof utilizes the result from McFadden and Train (2000) with the explanation from Train (2009). Since the utility from Assumption 2.3 is $U_{nj} = X_{nj}\beta + \epsilon_{nj}$ with choice probability P_{nj} then $P_{nj} = \lim_{c\to 0} \int \frac{e^{\frac{1}{c}(X_{nj}\beta+\epsilon_{nj})}}{\sum_{i=1}^{J} e^{\frac{1}{c}(X_{ni}\beta+\epsilon_{nj})}} dF_\epsilon$ (McFadden and Train, 2000). Then for some arbitrarily small c, $P_{nj} \approx \int \frac{e^{\frac{1}{c}(X_{nj}\beta+\epsilon_{nj})}}{\sum_{i=1}^{J} e^{\frac{1}{c}(X_{ni}\beta+\epsilon_{nj})}} dF_\epsilon$. But for any known c, the estimation results in the same estimates, thus it is equivalent to perform estimation as if $c = 1$. I then write $\epsilon_{nj} = Z_{nj}b_n$ and $F_\epsilon = F_b$ to conform to commonly accepted notation. □

The previous lemma was written in the notation of Assumption 2.2. However it holds equally for Assumption 2.3 where P_{nj}, β, Z_{nj}, b_n and F_b are replaced with P^0_{nj}, β^0, Z^0_{nj}, b^0_n and F^0_b. The next lemma provides a simple solution to the derivative of the integral in Lemma 2.1.

Lemma B.2. *Let* $P_{nj}|b_n = \frac{e^{X_{nj}\beta+Z_{nj}b_n}}{\sum_{i=1}^{J} e^{X_{ni}\beta+Z_{ni}b_n}}$ *and* $b_n \sim F_b$ *the CDF of a finite dimensional random vector, then* $\frac{d}{dX_{nr}} P_{nj} = \frac{d}{dX_{nr}} \int P_{nj}|b_n \, dF_b = \int \frac{d}{dX_{nr}} P_{nj}|b_n \, dF_b$ *for any* $r \in \{1, 2, ..., J\}$

Proof. Notice that dominated convergence holds since $0 \leq P_{nj}|b_n \leq 1$ and $\int 1F_b = 1 < \infty$. □

Lemma B.2 has additional useful implications for numerical optimization. Passing the derivative inside the integral improves computational time and accuracy. In fact the second derivative can be passed inside as well.[2] This result has never been proven before. The proof of

[2]

Theorem B.1. *Let* $P_{nj}|b_n = \frac{1}{\sum_{i=1}^{J} e^{(X_{ni}-X_{nj})\beta+(Z_{ni}-Z_{nj})b_n}}$ *and* $b_n \sim F_b$ *the CDF of a finite dimensional random vector. Suppose* $J < \infty$. *Then* $\frac{d}{d\beta^T} \int \frac{d}{d\beta} P_{nj}|b_n \, dF_b = \int \frac{d^2}{d\beta d\beta^T} P_{nj}|b_n \, dF_b$

Proof. Note $\frac{d}{d\beta_k} P_{nj}|b_n = -\frac{1}{(\sum_{l=1}^{J} e^{(X_{nl}-X_{nj})\beta+(Z_{nl}-Z_{nj})b_n})^2} \sum_{i=1}^{J} (X_{ni} - X_{nj})^{(k)} e^{(X_{ni}-X_{nj})\beta+(Z_{ni}-Z_{nj})b_n} =$ $-P_{nj}|b_n \sum_{i=1}^{J} \frac{(X_{ni}-X_{nj})^{(k)} e^{(X_{ni}-X_{nj})\beta+(Z_{ni}-Z_{nj})b_n}}{\sum_{l=1}^{J} e^{(X_{nl}-X_{nj})\beta+(Z_{nl}-Z_{nj})b_n}} = -P_{nj}|b_n \sum_{i=1}^{J} (X_{ni} - X_{nj})^{(k)} P_{ni}|b_n = -\sum_{i=1}^{J} (X_{ni} - X_{nj})^{(k)} P_{ni}|b_n P_{nj}|b_n$. Then $\int \frac{d}{d\beta} P_{nj}|b_n \, dF_b = \int -\sum_{i=1}^{J} (X_{ni} - X_{nj})^{(k)} P_{ni}|b_n P_{nj}|b_n \, dF_b = -\sum_{i=1}^{J} (X_{ni} - X_{nj})^{(k)} \int P_{ni}|b_n P_{nj}|b_n \, dF_b$. Since $0 \leq P_{ni}|b_n P_{nj}|b_n \leq 1$ and $\int 1 \, dF_b = 1 < \infty$, then dominated convergence holds. Thus the second derivative can be passed through the integral as well. □

Theorem 2.2 is below.

Proof. This proof is identical to that of Theorem 2.1. Using Lemmas 2.1 and B.1 the assumed choice probability is replaced with the mixed logit approximation. □

Lemma B.3. *Suppose that* $P_{nj} = \frac{e^{X_{nj}\beta}}{\sum_{i=1}^{J} e^{X_{ni}\beta}}$ *then* $\frac{dP_{nj}}{dX_{nj}} = P_{nj}(1-P_{nj})\beta$ *and* $\frac{dP_{nj}}{dX_{ni}} = -P_{nj}P_{ni}\beta$.

Proof. See Train (2009) page 58. □

Lemma B.4. *Suppose that* $P^0_{nj} = \int \frac{e^{X_{nj}\beta^0 + Z_{nj}b_n}}{\sum_{i=1}^{J} e^{X_{ni}\beta^0 + Z_{ni}b_n}} dFb$ *then* $\frac{dP^0_{nj}}{dX_{nj}} = P^0_{nj}(1 - P^0_{nj})\beta^0$ *and* $\frac{dP^0_{nj}}{dX_{ni}} = -P^0_{nj}P^0_{ni}\beta^0$.

Proof. The result follows from Lemmas B.1 and B.3. □

The proof of Theorem 2.3 is below.

Proof. I prove it for the case $sign(\beta^\star_1) = sign(\beta^0_1)$, the other cases follow by analogy. Without loss of generality, suppose $X^{(1)}_{11} \geq X^{(1)}_{1j}\ \forall j > 1$ (this can always be performed by reordering the alternatives). Evaluating $W(\beta)$ at the KL minimizer implies that $\sum_{n=1}^{N}\sum_{j=1}^{J} P_{nj}X^{(1)}_{nj} = \sum_{n=1}^{N}\sum_{j=1}^{J} P^0_{nj}X^{(1)}_{nj}$. Then $\frac{d}{dX^{(1)}_{11}}\sum_{n=1}^{N}\sum_{j=1}^{J} P_{nj}X^{(1)}_{nj}$

$$
\begin{aligned}
&= \sum_{j=1}^{J} \frac{d}{dX^{(1)}_{11}} P_{1j}X^{(1)}_{1j} \text{ (I now drop the individual subscript)} \\
&= P_1(1-P_1)X^{(1)}_1\beta^\star_1 + P_1 - P_1P_2X^{(1)}_2\beta^\star_1 - \ldots - P_1P_JX^{(1)}_J\beta^\star_1 \text{ by lemma B.3} \\
&= P_1 + \beta^\star_1 P_1((1-P_1)X^{(1)}_1 - P_2X^{(1)}_2 - \ldots - P_JX^{(1)}_J) \\
&= P_1 + \beta^\star_1 P_1((1-(1-P_2-\ldots-P_J))X^{(1)}_1 - P_2X^{(1)}_2 - \ldots - P_JX^{(1)}_J) \\
&= P_1 + \beta^\star_1 P_1(P_2(X^{(1)}_1 - X^{(1)}_2) + \ldots + P_J(X^{(1)}_1 - X^{(1)}_J)) \\
&\equiv P_1 + \beta^\star_1 Q
\end{aligned}
$$

Likewise $\frac{d}{dX_{11}^{(1)}} \sum_{n=1}^{N} \sum_{j=1}^{J} P_{nj}^0 X_{nj}^{(1)} = P_1^0 + \beta_1^0 P_1^0 (P_2^0 (X_1^{(1)} - X_2^{(1)}) + \ldots + P_J^0 (X_1^{(1)} - X_J^{(1)})) \equiv P_1^0 + \beta_1^0 Q^0$ by Lemmas 2.1, B.1 and B.3. Note that $Q > 0$ and $Q^0 > 0$ by Assumption 2.4.

Note that in all random utilty models, rescaling the covariates does not change the choice probability. Thus I can write $W(\beta)$ evaluated at the KL minimizer as $\sum_{n=1}^{N} \sum_{j=1}^{J} P_{nj} c X_{nj}^{(1)} = \sum_{n=1}^{N} \sum_{j=1}^{J} P_{nj}^0 c X_{nj}^{(1)}$, which must hold for all c. Which implies $\frac{d}{dX_{11}^{(1)}} \sum_{n=1}^{N} \sum_{j=1}^{J} P_{nj} X_{nj}^{(1)} = \frac{d}{dc X_{11}^{(1)}} \sum_{n=1}^{N} \sum_{j=1}^{J} P_{nj} c X_{nj}^{(1)} = P_1 + \beta_1^\star c Q$. Likewise $\frac{d}{dc X_{11}^{(1)}} \sum_{n=1}^{N} \sum_{j=1}^{J} P_{nj}^0 X_{nj}^{(1)} = P_1^0 + \beta_1^0 c Q^0$. And thus $P_1 + \beta_1^\star c Q = P_1^0 + \beta_1^0 c Q^0 \iff 0 = P_1^0 - P_1 + \beta_1^0 c Q^0 - \beta_1^\star c Q$.

For the sake of contradiction assume $sign(\beta^\star) \neq sign(\beta^0)$, then $sign(\beta_1^0 c Q^0) \neq sign(\beta_1^\star c Q)$ must hold for all $c \neq 0$, $\beta_1^0 \neq 0$ and $\beta_1^\star \neq 0$. Suppose $\beta^\star \geq 0$ and $\beta^0 \leq 0$ where one equality is strict. Also suppose $c > 0$. Then $\beta_1^0 c Q^0 - \beta_1^\star c Q < 0$ and $P_1^0 - P_1 > 0$. Then for a given $\beta_1^\star$ and β_1^0 there exists a $c^\dagger$ such that $\beta_1^0 c^\dagger Q^0 - \beta_1^\star c^\dagger Q < -2$. Note that $-1 \leq P_1^0 - P_1 \leq 1$. Then $0 = P_1^0 - P_1 + \beta_1^0 c^\dagger Q^0 - \beta_1^\star c^\dagger Q < -2 + 1 = -1$, a contradiction. Thus $sign(\beta_1^\star) = sign(\beta_1^0)$. By analogy this is true for all elements in $\beta^\star$ and β^0, thus $sign(\beta^\star) = sign(\beta^0)$. □

The proof of Corollary 2.1 is below.

Proof. From the proof of Theorem 2.3, since $\beta_k^0 = 0$ then $0 = P_1^0 - P_1 - \beta_1^\star c Q$. so $\beta_k^\star c Q = P_1^0 - P_1$. Since $-1 < P_1^0 - P_1 < 1$ then $-1 < \beta_k^\star c Q < 1$. Without loss of generality assume $c > 0$ then $-\frac{1}{cQ} < \beta_k^\star < \frac{1}{cQ}$ which must hold for any choice of c. Since $Q > 0$, it follows that $\beta_k^\star = 0$ by squeeze theorem. □

The proof of Corollary 2.2 is below

Proof. The result follows immediately from Theorem 2.3 and Corollary 2.2. The type one error rate being conservative follows from White (1983). □

B.2 Type one error rate

Corollary 2.2 states type one error rates for null coefficients will be at least asymptotically conservative. Tables B.1 and B.2 illustrate this result. The coefficients for the observable utility in the DGP were $(\beta_1^0, \beta_2^0) = (0, 1)$. Then a mixed logit DGP (Table B.1) and a heteroskedastic logit DGP (Table B.2) were simulated but a conditional logit was estimated. The alpha level was set to $\alpha = 0.20$. The top row in Table B.1 shows the different variances in the random effect. The top row in Table B.2 shows the different levels of heteroskedasticity. The 'J' and 'N' columns represent the number of presented alternatives and individuals respectively. The values in the table are the simulated type one error rates for β_1.

		$\sigma^2 = 0.5^2$		$\sigma^2 = 1^2$		$\sigma^2 = 2^2$	
J	N	H	R	H	R	H	R
2	100	0.18	0.18	0.18	0.15	0.15	0.10
2	500	0.19	0.20	0.18	0.19	0.15	0.16
2	1000	0.18	0.18	0.15	0.16	0.10	0.10
3	100	0.20	0.20	0.22	0.22	0.24	0.24
3	500	0.18	0.18	0.18	0.18	0.13	0.13
3	1000	0.19	0.19	0.19	0.20	0.16	0.16
5	100	0.18	0.20	0.19	0.21	0.20	0.22
5	500	0.19	0.19	0.18	0.19	0.14	0.14
5	1000	0.19	0.20	0.19	0.20	0.17	0.19

H: Hessian standard error

R: Huber-White robust standard error

Table B.1: Type one error rate for $\alpha = 0.20$ (mixed logit DGP)

		$\gamma_1 = 0.5$		$\gamma_1 = 1$		$\gamma_1 = 1.5$	
J	N	H	R	H	R	H	R
2	100	0.18	0.21	0.18	0.20	0.18	0.20
2	500	0.19	0.18	0.20	0.19	0.19	0.18
2	1000	0.20	0.19	0.20	0.20	0.21	0.21
3	100	0.18	0.19	0.19	0.21	0.20	0.23
3	500	0.18	0.18	0.19	0.18	0.22	0.21
3	1000	0.18	0.18	0.18	0.18	0.19	0.19
5	100	0.22	0.24	0.18	0.19	0.19	0.21
5	500	0.21	0.20	0.21	0.21	0.20	0.20
5	1000	0.20	0.20	0.20	0.19	0.20	0.20

H: Hessian standard error

R: Huber-White robust standard error

Table B.2: Type one error rate for $\alpha = 0.20$ (heteroskedastic logit DGP)

B.3 Kullback-Leibler minimizer estimand

When the model is misspecified the QML estimator is estimating the parameter minimizing the KL divergence from the assumed model to the true model. Tables B.3 and B.4 show the KL minimizer from the simulation failing to specificy a random effect (Table B.3) and heteroskedastic effect (Table B.4). The top row in Table B.3 shows the different variances in the random effect. The top row in Table B.4 shows the different levels of heteroskedasticity. The 'J' and 'N' columns represent the number of presented alternatives and individuals respectively. The β column identifies the parameter of interest. The coefficients for the observable utility in the DGP were $(\beta_1^0, \beta_2^0) = (-2, 1)$. The '$\beta^\star$' column shows the analytic KL minimizer found by using numerical optimization. The 'SE' column shows the estimated standard error of the QML estimates around the KL minimizer over the simulation. This standard error is calculated by taking the square root of the sample variance of the estimated beta coefficients around the analytic KL minimizer, $\sqrt{S^{-1}\sum_{s=1}^{S}(\hat{\beta}_{is} - \beta_i^\star)^2}$ for $i \in \{1, 2\}$ where S is the number of simulations. The 'KL' column shows the KL divergence evaluated at the KL minimizing parameter.

			$\sigma^2 = 0.5^2$			$\sigma^2 = 1^2$			$\sigma^2 = 2^2$		
J	N	β	$\beta^\star$	SE	KL	$\beta^\star$	SE	KL	$\beta^\star$	SE	KL
2	100	-2	-1.80	0.46	2.2	-1.48	0.35	7.8	-0.93	0.20	21.4
2	500	-2	-1.91	0.21	8.9	-1.67	0.17	31.7	-1.20	0.10	88.6
2	1000	-2	-1.88	0.14	16.4	-1.63	0.11	59.2	-1.17	0.07	171.7
3	100	-2	-1.93	0.40	2.6	-1.75	0.36	9.2	-1.30	0.25	25.5
3	500	-2	-1.97	0.17	10.3	-1.84	0.16	36.1	-1.50	0.12	100.9
3	1000	-2	-1.92	0.11	19.0	-1.77	0.11	68.0	-1.40	0.08	197.2
5	100	-2	-1.98	0.36	2.3	-1.90	0.37	8.1	-1.58	0.29	24.4
5	500	-2	-2.00	0.16	9.3	-1.96	0.15	34.0	-1.73	0.14	104.0
5	1000	-2	-1.97	0.11	16.7	-1.90	0.11	62.6	-1.65	0.09	197.7
2	100	1	0.82	0.50	2.2	0.64	0.41	7.8	0.44	0.31	21.4
2	500	1	0.89	0.22	8.9	0.76	0.19	31.7	0.54	0.14	88.6
2	1000	1	0.95	0.17	16.4	0.85	0.14	59.2	0.67	0.11	171.7
3	100	1	0.92	0.43	2.6	0.85	0.36	9.2	0.74	0.30	25.5
3	500	1	0.93	0.19	10.3	0.84	0.17	36.1	0.68	0.14	100.9
3	1000	1	0.97	0.13	19.0	0.92	0.12	68.0	0.80	0.11	197.2
5	100	1	1.01	0.38	2.3	1.02	0.38	8.1	0.99	0.34	24.4
5	500	1	0.98	0.16	9.3	0.93	0.16	34.0	0.81	0.14	104.0
5	1000	1	0.99	0.11	16.7	0.95	0.11	62.6	0.86	0.10	197.7

$\beta^\star$: KL minimizer, SE: Standard error of $\hat{\beta}$ around $\beta^\star$

KL: minimized KL distance of assumed model from DGP

Table B.3: KL minimizer (mixed logit DGP)

			$\gamma_1 = 0.5$			$\gamma_1 = 1$			$\gamma_1 = 1.5$		
J	N	β	$\beta^\star$	SE	KL	$\beta^\star$	SE	KL	$\beta^\star$	SE	KL
2	100	-2	-1.67	0.49	4.2	-1.24	0.40	10.5	-1.03	0.33	14.9
2	500	-2	-1.60	0.20	21.7	-1.07	0.14	58.1	-0.80	0.11	84.9
2	1000	-2	-1.64	0.14	41.4	-1.13	0.11	112.9	-0.86	0.09	168.1
3	100	-2	-1.76	0.41	6.4	-1.35	0.35	16.7	-1.10	0.31	24.3
3	500	-2	-1.66	0.17	31.1	-1.15	0.14	84.9	-0.87	0.11	127.5
3	1000	-2	-1.70	0.12	61.6	-1.21	0.10	169.7	-0.92	0.08	254.8
5	100	-2	-1.81	0.35	7.0	-1.47	0.34	19.7	-1.23	0.31	31.2
5	500	-2	-1.73	0.16	39.2	-1.27	0.14	113.7	-0.98	0.13	177.9
5	1000	-2	-1.73	0.11	80.3	-1.27	0.10	230.2	-0.97	0.09	353.9
2	100	1	0.71	0.55	4.2	0.51	0.48	10.5	0.41	0.44	14.9
2	500	1	0.96	0.21	21.7	0.84	0.18	58.1	0.75	0.16	84.9
2	1000	1	0.90	0.15	41.4	0.74	0.13	112.9	0.63	0.12	168.1
3	100	1	0.94	0.38	6.4	0.76	0.35	16.7	0.65	0.30	24.3
3	500	1	1.01	0.18	31.1	0.93	0.16	84.9	0.87	0.15	127.5
3	1000	1	1.01	0.13	61.6	0.91	0.12	169.7	0.83	0.10	254.8
5	100	1	0.89	0.34	7.0	0.74	0.31	19.7	0.64	0.28	31.2
5	500	1	0.99	0.16	39.2	0.92	0.15	113.7	0.87	0.14	177.9
5	1000	1	1.02	0.11	80.3	0.97	0.11	230.2	0.92	0.10	353.9

$\beta^\star$: KL minimizer, SE: Standard error of $\hat{\beta}$ around $\beta^\star$

KL: minimized KL distance of assumed model from DGP

Table B.4: KL minimizer (heteroskedastic logit DGP)

B.4 Coverage probabilities

Coverage probabilities of the QML estimators for the data generating parameters, β^0, are presented in Tables B.5, B.6, and B.7. Coverage probabilities of the QML estimators for the KL minimizer parameters, $\beta^\star$, are presented in Tables B.8 and B.9. The column J represents the number of alternatives and N represents the number of individuals. The coefficients for the observable utility in the data generating process were $(\beta_1^0, \beta_2^0) = (-2, 1)$. Three types of standard errors are used: Hessian (denoted by H), Huber-White robust (denoted by R), and simulation (denoted by S). The simulation standard error is calculated by taking the square root of the variance of the estimated β coefficients around the data generating coefficient, $\sqrt{S^{-1}\sum_{s=1}^{S}(\hat{\beta}_{is} - \beta_i^0)^2}$ for $i \in \{1, 2\}$ where S is the number of simulations. The simulation based standard errors help show if normality of the estimator is being achieved. The confidence intervals calculated have level 80%, thus a better performing estimator will have a coverage probability closer to 0.80.

Table B.5 shows the coverage probabilities of the correctly specified conditional logit. Table B.6 and B.8 shows the coverage probabilities of the misspecified conditional logit failing to account for individual level heteroskedasticity. Coverage probabilities for the misspecified conditional logit (denoted M) and the correctly specified heteroskedastic logit (denoted C) are given. In the correctly specified heteroskedastic logit the heteroskedastic parameter θ_n is estimated (but omitted from the table). Table B.7 and B.9 shows the coverage probabilities for the misspecified conditional logit failing to account for a random alternative specific effect. Coverage probabilities for the misspecified conditional logit (denoted M) and the correctly specified mixed logit (denoted C) are given.

J	I	β	H	R	S
2	100	β_1	0.80	0.79	0.86
2	500	β_1	0.80	0.79	0.81
2	1000	β_1	0.82	0.82	0.81
3	100	β_1	0.80	0.79	0.84
3	500	β_1	0.77	0.77	0.80
3	1000	β_1	0.79	0.78	0.80
5	100	β_1	0.83	0.81	0.82
5	500	β_1	0.78	0.78	0.81
5	1000	β_1	0.79	0.79	0.80
2	100	β_2	0.82	0.80	0.82
2	500	β_2	0.82	0.82	0.82
2	1000	β_2	0.80	0.81	0.82
3	100	β_2	0.80	0.79	0.82
3	500	β_2	0.79	0.80	0.80
3	1000	β_2	0.80	0.80	0.80
5	100	β_2	0.80	0.80	0.82
5	500	β_2	0.82	0.81	0.80
5	1000	β_2	0.80	0.80	0.80

H: Hessian standard error

R: Huber-White robust standard error

S: Simulation standard error

Table B.5: Coverage probabilities of DGP parameters (conditional logit correctly specified)

			$\sigma^2 = 0.5^2$					$\sigma^2 = 1^2$					$\sigma^2 = 2^2$				
J	I	β	HM	RM	HC	RC	SC	HM	RM	HC	RC	SC	HM	RM	HC	RC	SC
2	100	β_1	0.76	0.74	0.98	0.98	0.99	0.48	0.46	0.96	0.96	0.99	0.00	0.00	0.90	0.90	0.98
2	500	β_1	0.77	0.76	0.96	0.97	0.79	0.33	0.32	0.91	0.91	0.82	0.00	0.00	0.65	0.65	0.87
2	1000	β_1	0.66	0.65	0.96	0.96	0.86	0.04	0.04	0.83	0.83	0.88	0.00	0.00	0.63	0.63	0.86
3	100	β_1	0.80	0.79	0.61	0.61	0.89	0.62	0.63	0.62	0.62	0.90	0.14	0.16	0.62	0.62	0.90
3	500	β_1	0.81	0.81	0.39	0.39	0.68	0.59	0.58	0.39	0.39	0.68	0.01	0.01	0.44	0.44	0.73
3	1000	β_1	0.74	0.73	0.42	0.42	0.78	0.23	0.22	0.40	0.40	0.78	0.00	0.00	0.39	0.39	0.75
5	100	β_1	0.80	0.80	0.55	0.55	0.68	0.77	0.77	0.60	0.60	0.73	0.44	0.47	0.66	0.67	0.79
5	500	β_1	0.79	0.79	0.63	0.63	0.81	0.79	0.80	0.66	0.66	0.86	0.29	0.31	0.61	0.61	0.85
5	1000	β_1	0.78	0.79	0.47	0.47	0.61	0.62	0.62	0.43	0.43	0.61	0.01	0.01	0.34	0.34	0.67
2	100	β_2	0.81	0.79	0.96	0.96	0.98	0.75	0.72	0.96	0.96	0.99	0.51	0.44	0.92	0.93	0.98
2	500	β_2	0.77	0.76	0.96	0.96	0.83	0.58	0.56	0.92	0.92	0.87	0.09	0.08	0.79	0.79	0.82
2	1000	β_2	0.78	0.77	0.97	0.97	0.98	0.66	0.65	0.96	0.96	0.95	0.10	0.09	0.88	0.89	0.90
3	100	β_2	0.80	0.79	0.73	0.74	0.88	0.81	0.79	0.75	0.76	0.90	0.74	0.73	0.75	0.76	0.86
3	500	β_2	0.76	0.76	0.48	0.48	0.89	0.63	0.62	0.50	0.51	0.89	0.25	0.24	0.51	0.51	0.87
3	1000	β_2	0.79	0.79	0.53	0.53	0.86	0.75	0.75	0.50	0.50	0.86	0.37	0.36	0.47	0.47	0.87
5	100	β_2	0.80	0.81	0.71	0.71	0.83	0.80	0.80	0.73	0.74	0.84	0.84	0.84	0.73	0.74	0.83
5	500	β_2	0.82	0.82	0.64	0.64	0.79	0.75	0.75	0.65	0.65	0.83	0.54	0.54	0.59	0.59	0.87
5	1000	β_2	0.80	0.80	0.47	0.47	0.73	0.78	0.78	0.43	0.43	0.71	0.52	0.52	0.41	0.41	0.86

First Letter H: Hessian standard error, R: Huber-White robust standard error, S: Simulation standard error

Second Letter M: Misspecified model, C: Correct model

Table B.6: Coverage probabilities of DGP parameters (mixed logit DGP)

			$\gamma_1 = 0.5$					$\gamma_1 = 1$					$\gamma_1 = 1.5$				
J	I	β	HM	RM	HC	RC	SC	HM	RM	HC	RC	SC	HM	RM	HC	RC	SC
2	100	β_1	0.64	0.69	0.79	0.74	0.88	0.24	0.30	0.80	0.75	1.00	0.08	0.12	0.81	0.77	0.99
2	500	β_1	0.22	0.28	0.78	0.78	0.82	0.00	0.00	0.80	0.79	0.81	0.00	0.00	0.81	0.80	0.82
2	1000	β_1	0.09	0.12	0.80	0.79	0.81	0.00	0.00	0.80	0.79	0.80	0.00	0.00	0.78	0.78	0.81
3	100	β_1	0.67	0.73	0.79	0.77	0.85	0.26	0.33	0.84	0.81	0.86	0.09	0.13	0.80	0.77	0.87
3	500	β_1	0.22	0.28	0.81	0.80	0.80	0.00	0.00	0.82	0.81	0.82	0.00	0.00	0.78	0.78	0.81
3	1000	β_1	0.10	0.14	0.80	0.80	0.81	0.00	0.00	0.81	0.81	0.81	0.00	0.00	0.78	0.78	0.80
5	100	β_1	0.72	0.76	0.81	0.79	0.81	0.34	0.44	0.77	0.76	0.82	0.12	0.17	0.84	0.81	0.83
5	500	β_1	0.31	0.39	0.81	0.80	0.81	0.00	0.00	0.80	0.79	0.80	0.00	0.00	0.82	0.80	0.82
5	1000	β_1	0.12	0.17	0.79	0.79	0.81	0.00	0.00	0.79	0.79	0.80	0.00	0.00	0.79	0.79	0.79
2	100	β_2	0.74	0.73	0.80	0.74	0.85	0.59	0.60	0.80	0.67	0.95	0.51	0.51	0.85	0.64	0.96
2	500	β_2	0.82	0.82	0.80	0.80	0.80	0.68	0.68	0.82	0.79	0.82	0.47	0.46	0.83	0.79	0.82
2	1000	β_2	0.73	0.73	0.79	0.78	0.80	0.28	0.28	0.81	0.80	0.80	0.06	0.06	0.80	0.76	0.81
3	100	β_2	0.83	0.82	0.78	0.74	0.84	0.74	0.73	0.82	0.71	0.86	0.66	0.64	0.85	0.75	0.89
3	500	β_2	0.83	0.82	0.79	0.78	0.81	0.80	0.78	0.80	0.77	0.82	0.70	0.70	0.81	0.78	0.80
3	1000	β_2	0.81	0.81	0.79	0.79	0.80	0.71	0.70	0.79	0.77	0.81	0.46	0.45	0.80	0.78	0.78
5	100	β_2	0.81	0.81	0.79	0.77	0.83	0.71	0.69	0.79	0.74	0.82	0.57	0.54	0.88	0.78	0.83
5	500	β_2	0.82	0.82	0.81	0.81	0.80	0.76	0.76	0.80	0.78	0.81	0.66	0.66	0.82	0.77	0.81
5	1000	β_2	0.82	0.83	0.81	0.80	0.79	0.80	0.80	0.77	0.76	0.79	0.70	0.71	0.80	0.77	0.81

First Letter H: Hessian standard error, R: Huber-White robust standard error, S: Simulation standard error

Second Letter M: Misspecified model, C: Correct model

Table B.7: Coverage probabilities of DGP parameters (heteroskedastic logit DGP)

			$\sigma^2 = 0.5^2$			$\sigma^2 = 1^2$			$\sigma^2 = 2^2$		
J	N	β	H	R	S	H	R	S	H	R	S
2	100	β_1	0.82	0.81	0.84	0.84	0.82	0.82	0.90	0.89	0.83
2	500	β_1	0.81	0.80	0.83	0.84	0.83	0.81	0.91	0.88	0.80
2	1000	β_1	0.82	0.80	0.82	0.85	0.83	0.81	0.92	0.89	0.83
3	100	β_1	0.81	0.80	0.82	0.80	0.81	0.84	0.85	0.85	0.81
3	500	β_1	0.82	0.82	0.80	0.81	0.82	0.81	0.86	0.85	0.79
3	1000	β_1	0.82	0.81	0.79	0.80	0.80	0.81	0.86	0.85	0.80
5	100	β_1	0.80	0.80	0.81	0.79	0.78	0.84	0.82	0.84	0.82
5	500	β_1	0.79	0.79	0.79	0.82	0.82	0.80	0.82	0.83	0.82
5	1000	β_1	0.80	0.80	0.80	0.81	0.81	0.82	0.83	0.83	0.81
2	100	β_2	0.84	0.80	0.78	0.91	0.87	0.79	0.94	0.90	0.84
2	500	β_2	0.82	0.82	0.80	0.87	0.86	0.79	0.93	0.92	0.78
2	1000	β_2	0.81	0.80	0.81	0.86	0.85	0.79	0.91	0.90	0.81
3	100	β_2	0.81	0.80	0.83	0.85	0.84	0.81	0.89	0.88	0.82
3	500	β_2	0.80	0.80	0.81	0.81	0.81	0.79	0.88	0.87	0.78
3	1000	β_2	0.80	0.79	0.79	0.82	0.82	0.78	0.87	0.86	0.81
5	100	β_2	0.80	0.80	0.81	0.80	0.80	0.82	0.83	0.84	0.80
5	500	β_2	0.82	0.82	0.81	0.81	0.81	0.80	0.85	0.85	0.80
5	1000	β_2	0.80	0.80	0.77	0.81	0.81	0.78	0.86	0.86	0.80

H: Hessian standard error, R: Huber White robust standard error

S: Simulation standard error

Table B.8: Coverage probabilities of KL minimizer (mixed logit DGP)

			$\gamma_1 = 0.5$			$\gamma_1 = 1$			$\gamma_1 = 1.5$		
J	N	β	H	R	S	H	R	S	H	R	S
2	100	β_1	0.77	0.80	0.84	0.75	0.82	0.87	0.79	0.88	0.86
2	500	β_1	0.74	0.81	0.82	0.73	0.85	0.80	0.74	0.88	0.81
2	1000	β_1	0.76	0.82	0.81	0.72	0.84	0.83	0.72	0.87	0.83
3	100	β_1	0.77	0.81	0.83	0.75	0.85	0.83	0.74	0.84	0.84
3	500	β_1	0.75	0.82	0.80	0.72	0.85	0.82	0.70	0.86	0.81
3	1000	β_1	0.75	0.82	0.81	0.68	0.84	0.79	0.71	0.88	0.79
5	100	β_1	0.78	0.82	0.82	0.72	0.82	0.82	0.73	0.85	0.83
5	500	β_1	0.75	0.83	0.81	0.70	0.85	0.81	0.66	0.84	0.81
5	1000	β_1	0.75	0.84	0.80	0.69	0.85	0.79	0.68	0.86	0.84
2	100	β_2	0.82	0.82	0.82	0.79	0.82	0.78	0.85	0.84	0.82
2	500	β_2	0.83	0.83	0.80	0.86	0.86	0.80	0.87	0.87	0.80
2	1000	β_2	0.84	0.83	0.80	0.86	0.86	0.80	0.87	0.88	0.82
3	100	β_2	0.83	0.82	0.80	0.85	0.83	0.80	0.90	0.89	0.81
3	500	β_2	0.82	0.82	0.82	0.86	0.85	0.81	0.88	0.87	0.82
3	1000	β_2	0.81	0.81	0.80	0.85	0.85	0.83	0.88	0.87	0.79
5	100	β_2	0.83	0.82	0.81	0.85	0.84	0.81	0.87	0.86	0.81
5	500	β_2	0.82	0.82	0.80	0.85	0.84	0.82	0.85	0.85	0.81
5	1000	β_2	0.83	0.83	0.80	0.82	0.83	0.80	0.84	0.85	0.79

H: Hessian standard error, R: Huber White robust standard error

S: Simulation standard error

Table B.9: Coverage probabilities of KL minimizer (heteroskedastic logit DGP)

B.5 MSE of choice probabilities

Square root of MSE of choice probabilities is presented in Tables B.10 and B.11. In Table B.10 the data generating process is mixed logit; a conditional logit and a mixed logit are estimated. In Table B.11 the data generating process is heteroskedastic logit; a conditional logit and a heteroskedastic logit are estimated. The square root MSE is calculated as $S^{-1}\sum_{s=1}^{S}\sqrt{N^{-1}\sum_{n=1}^{N}\sum_{j=1}^{J}(\hat{P}_{njs}-P^0_{njs})^2}$, where $\hat{P}_{njs}$ is the estimated choice probability and P^0_{njs} is the true choice probability for individual n with alternative j in simulation s. The choice probabilities for the mixed logit are with the individual effects marginalized out. The choice probabilities have been scaled by 10 for presentation purposes.

		$\sigma^2=0.5^2$		$\sigma^2=1^2$		$\sigma^2=2^2$	
J	N	M	C	M	C	M	C
2	100	0.46	1.08	0.46	1.01	0.56	1.00
2	500	0.22	0.62	0.22	0.61	0.23	0.55
2	1000	0.16	0.54	0.17	0.49	0.21	0.55
3	100	0.38	1.24	0.39	1.21	0.52	1.07
3	500	0.18	1.46	0.22	1.37	0.39	1.17
3	1000	0.13	1.54	0.19	1.51	0.38	1.24
5	100	0.25	0.59	0.29	0.56	0.45	0.53
5	500	0.12	0.36	0.18	0.33	0.38	0.48
5	1000	0.09	0.57	0.16	0.57	0.37	0.60

M: Misspecified conditional logit

C: Correctly specified mixed logit

Table B.10: Square root MSE of choice probabilities (mixed logit DGP)

		$\gamma_1 = 0.5$		$\gamma_1 = 1$		$\gamma_1 = 1.5$	
J	N	M	C	M	C	M	C
2	100	1.09	1.55	1.73	1.81	2.12	2.18
2	500	0.99	1.28	1.73	1.65	2.20	2.12
2	1000	0.97	1.28	1.70	1.63	2.17	2.08
3	100	1.04	1.38	1.65	1.61	2.05	2.03
3	500	0.96	1.25	1.64	1.56	2.09	2.01
3	1000	0.94	1.25	1.62	1.54	2.07	1.98
5	100	0.74	1.07	1.26	1.24	1.65	1.64
5	500	0.76	1.02	1.35	1.27	1.76	1.71
5	1000	0.77	1.03	1.35	1.28	1.74	1.69

M: Misspecified conditional logit

C: Correctly specified heteroskedastic logit

Table B.11: Square root MSE of choice probabilities (heteroskedastic logit DGP)

Appendix C

Chapter 3

C.1 Lemmas and proofs

In this section I prove the posterior is consistent for the parameters of interest. This proof is for the location case. The regression case should come with easy modification by concatenating $\beta_{\boldsymbol{\tau y}}$ with $\beta_{\boldsymbol{\tau x}}$ and $\mathbf{Y}^{\perp}_{\mathbf{u}i}$ with $\mathbf{X}_i$. Since $\mathbf{Y}$ and $\mathbf{X}$ rely on the same sets of assumptions and also expectations are taken over $\mathbf{Y}$ and $\mathbf{X}$, there should not be any issue with these results generalizing to the regression case.

Define the population parameters $(\alpha_{\boldsymbol{\tau}0}, \beta_{\boldsymbol{\tau}0})$ to be the parameters that satisfy (3.11) and (3.12). Note that the posterior can be written equivalently as

$$\Pi_{\boldsymbol{\tau}}(U|(\mathbf{Y}_1,\mathbf{X}_1),(\mathbf{Y}_2,\mathbf{X}_2),...,(\mathbf{Y}_n,\mathbf{X}_n)) = \frac{\int_U \prod_{i=1}^n \frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\mathbf{X}_i,\alpha_{\boldsymbol{\tau}},\beta_{\boldsymbol{\tau}},\sigma_{\boldsymbol{\tau}})}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\mathbf{X}_i,\alpha_{\boldsymbol{\tau}0},\beta_{\boldsymbol{\tau}0},\sigma_{\boldsymbol{\tau}0})} d\Pi_{\boldsymbol{\tau}}(\alpha_{\boldsymbol{\tau}},\beta_{\boldsymbol{\tau}})}{\int_\Theta \prod_{i=1}^n \frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\mathbf{X}_i\alpha_{\boldsymbol{\tau}},\beta_{\boldsymbol{\tau}},\sigma_{\boldsymbol{\tau}})}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\mathbf{X}_i,\alpha_{\boldsymbol{\tau}0},\beta_{\boldsymbol{\tau}0},\sigma_{\boldsymbol{\tau}0})} d\Pi_{\boldsymbol{\tau}}(\alpha_{\boldsymbol{\tau}},\beta_{\boldsymbol{\tau}})} \quad \text{(C.1)}$$

For ease of readability I will omit $\boldsymbol{\tau}$ from $\alpha_{\boldsymbol{\tau}}, \beta_{\boldsymbol{\tau}}$ and $\Pi_{\boldsymbol{\tau}}$. Writing the posterior in this form

is for mathematical convenience. It allows me to focus on the numerator,

$$I_n(U) = \int_U \prod_{i=1}^{n} \frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,\sigma)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,\sigma)} d\Pi(\alpha,\beta), \tag{C.2}$$

and denominator, $I_n(\Theta)$, separately. The next lemma provides several inequalities that are useful later and is presented without proof.

Lemma C.1. *Let* $b_i = (\alpha-\alpha_0)+(\beta-\beta_0)'\mathbf{Y}_{\mathbf{u}i}^{\perp}$, $W_i = (\mathbf{u}'-\beta_0'\boldsymbol{\Gamma}_{\mathbf{u}}')\mathbf{Y}_i-\alpha_0$, $W_i^+ = max(W_i,0)$ *and* $W_i^- = min(-W_i,0)$. *Then a)* $\log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,\sigma)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,\sigma)}\right) =$

$$\frac{1}{\sigma}\begin{cases} -b_i(1-\tau) & \text{if } (\mathbf{u}'-\beta'\boldsymbol{\Gamma}_{\mathbf{u}}')\mathbf{Y}_i-\alpha\le 0 \text{ and } (\mathbf{u}'-\beta_0'\boldsymbol{\Gamma}_{\mathbf{u}}')\mathbf{Y}_i-\alpha_0\le 0 \\ -((\mathbf{u}'-\beta_0'\boldsymbol{\Gamma}_{\mathbf{u}}'')\mathbf{Y}_i-\alpha_0)+b_i\tau & \text{if } (\mathbf{u}'-\beta'\boldsymbol{\Gamma}_{\mathbf{u}}')\mathbf{Y}_i-\alpha> 0 \text{ and } (\mathbf{u}'-\beta_0'\boldsymbol{\Gamma}_{\mathbf{u}}')\mathbf{Y}_i-\alpha_0\le 0 \\ (\mathbf{u}'-\beta'\boldsymbol{\Gamma}_{\mathbf{u}}')\mathbf{Y}_i-\alpha+b_i\tau & \text{if } (\mathbf{u}'-\beta'\boldsymbol{\Gamma}_{\mathbf{u}}')\mathbf{Y}_i-\alpha\le 0 \text{ and } (\mathbf{u}'-\beta_0'\boldsymbol{\Gamma}_{\mathbf{u}}')\mathbf{Y}_i-\alpha_0> 0 \\ b_i\tau & \text{if } (\mathbf{u}'-\beta'\boldsymbol{\Gamma}_{\mathbf{u}}')\mathbf{Y}_i-\alpha> 0 \text{ and } (\mathbf{u}'-\beta_0'\boldsymbol{\Gamma}_{\mathbf{u}}')\mathbf{Y}_i-\alpha_0> 0 \end{cases}$$

b) $\log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,\sigma)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,\sigma)}\right) \le \frac{1}{\sigma}|b_i| \le |\alpha-\alpha_0| + |(\beta-\beta_0)'||\boldsymbol{\Gamma}_{\mathbf{u}}'||\mathbf{Y}_i|$

c) $\log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,\sigma)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,\sigma)}\right) \le \frac{1}{\sigma}|(\mathbf{u}'-\beta_0'\boldsymbol{\Gamma}_{\mathbf{u}}')\mathbf{Y}_i-\alpha_0| \le \frac{1}{\sigma}(|(\mathbf{u}'-\beta_0'\boldsymbol{\Gamma}_{\mathbf{u}}')||\mathbf{Y}_i|+|\alpha_0|)$

d) $\log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,\sigma)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,\sigma)}\right) = \frac{1}{\sigma}\begin{cases} -b_i(1-\tau)+\min(W_i^+,b_i) & \text{if } b_i>0 \\ b_i\tau+\min(W_i^-,-b_i) & \text{if } b_i\le 0 \end{cases}$

e) $\log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,\sigma)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,\sigma)}\right) \ge -\frac{1}{\sigma}|b_i| \ge -|\alpha-\alpha_0| - |(\beta-\beta_0)'||\boldsymbol{\Gamma}_{\mathbf{u}}'||\mathbf{Y}_i|$

The next lemma provides more useful inequalities.

Lemma C.2. *The following inequalities hold:*

a) $E\left[\log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,\sigma)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,\sigma)}\right)\right] \le 0$

b) $\sigma E\left[\log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,\sigma)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,\sigma)}\right)\right] = E\left[-(W_i-b_i)1_{(b_i<W_i<0)}\right] + E\left[(W_i-b_i)1_{(0<W_i<b_i)}\right]$

c) $\sigma E\left[\log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,\sigma)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,\sigma)}\right)\right] \leq E\left[-(W_i-b_i)\right]Pr(b_i<W_i<0) + E\left[(W_i-b_i)\right]Pr(0<W_i<b_i)$

d) $\sigma E\left[\log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,\sigma)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,\sigma)}\right)\right] \leq -E\left[-\frac{b_i}{2}1_{(b_i<0)}\right]Pr(\frac{b_i}{2}<W_i<0) - E\left[\frac{b_i}{2}1_{(0<b_i)}\right]Pr(0<W_i<\frac{b_i}{2})$

e) if Assumption 3.4 holds then $\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}E[|W_i|]<\infty$.

Proof. Note that $E[b_i] = (\alpha-\alpha_0)+(\beta-\beta_0)'E[\mathbf{Y}_{\mathbf{u}i}^{\perp}] = (\alpha-\alpha_0)+\frac{1}{\tau}(\beta-\beta_0)'E[\mathbf{Y}_{\mathbf{u}i}^{\perp}1_{(\mathbf{u}'-\beta_0'\boldsymbol{\Gamma}_{\mathbf{u}}')\mathbf{Y}_i-\alpha_0\leq 0)}]$ from subgradient condition (3.12). Define A_i to be the event $(\mathbf{u}'-\beta_0'\boldsymbol{\Gamma}_{\mathbf{u}}')\mathbf{Y}_i-\alpha_0\leq 0$ and A_i^c it's complement. Define B_i to be the event $(\mathbf{u}'-\beta'\boldsymbol{\Gamma}_{\mathbf{u}}')\mathbf{Y}_i-\alpha\leq 0$ and B_i^c it's complement.

$$
\begin{aligned}
&\sigma\log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,\sigma)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,\sigma)}\right)\\
&= b_i\tau - b_i1_{(A_i,B_i)} - ((\mathbf{u}'-\beta_0'\boldsymbol{\Gamma}_{\mathbf{u}}')\mathbf{Y}_i-\alpha_0)1_{(A_i,B_i^c)} + ((\mathbf{u}'-\beta'\boldsymbol{\Gamma}_{\mathbf{u}}')\mathbf{Y}_i-\alpha)1_{(A_i^c,B_i)}\\
&= b_i\tau - b_i1_{(A_i)} + (b_i-((\mathbf{u}'-\beta_0'\boldsymbol{\Gamma}_{\mathbf{u}}')\mathbf{Y}_i-\alpha_0))1_{(A_i,B_i^c)} + ((\mathbf{u}'-\beta'\boldsymbol{\Gamma}_{\mathbf{u}}')\mathbf{Y}_i-\alpha)1_{(A_i^c,B_i)}\\
&= b_i\tau - b_i1_{(A_i)} - ((\mathbf{u}'_i-\beta'\boldsymbol{\Gamma}_{\mathbf{u}}')\mathbf{Y}_i-\alpha)1_{(A_i,B_i^c)} + ((\mathbf{u}'-\beta'\boldsymbol{\Gamma}_{\mathbf{u}}')\mathbf{Y}_i-\alpha)1_{(A_i^c,B_i)}
\end{aligned}
$$

Since $E[(\alpha-\alpha_0)1_{(A_i)}] = \tau(\alpha-\alpha_0)$ then $E[b_i\tau - b_i1_{(A_i)}] = 0$. Then

$$\sigma E\left[\log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,\sigma)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,\sigma)}\right)\right] = E[-((\mathbf{u}'-\beta'\boldsymbol{\Gamma}_{\mathbf{u}}')\mathbf{Y}_i-\alpha)1_{(A_i,B_i^c)}]+E[(\mathbf{u}'-\beta'\boldsymbol{\Gamma}_{\mathbf{u}}')\mathbf{Y}_i-\alpha)1_{(A_i^c,B_i)}]$$

The constraint in the first term and second terms imply $-((\mathbf{u}'-\beta'\boldsymbol{\Gamma}_{\mathbf{u}}')\mathbf{Y}_i-\alpha)<0$ and $(\mathbf{u}'-\beta'\boldsymbol{\Gamma}_{\mathbf{u}}')\mathbf{Y}_i-\alpha\leq 0$ over their respective domains of integration. It follows

$$\sigma E\left[\log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,\sigma)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,\sigma)}\right)\right] = E\left[-(W_i-b_i)1_{(b_i<W_i<0)}\right] + E\left[(W_i-b_i)1_{(0<W_i<b_i)}\right].$$

Note that $(W_i - b_i)1_{(0<W_i<b_i)} \leq (W_i - b_i)1_{(0<W_i<\frac{b_i}{2})} < -\frac{b_i}{2}1_{(0<W_i<\frac{b_i}{2})}$. Likewise, $-(W_i - b_i)1_{(b_i<W_i<0)} < \frac{b_i}{2}1_{(\frac{b_i}{2}<W_i<0)}$. Thus,

$$\sigma E\left[\log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,\sigma)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,\sigma)}\right)\right] \leq E\left[\frac{b_i}{2}1_{(\frac{b_i}{2}<W_i<0)}\right] + E\left[-\frac{b_i}{2}1_{(\frac{b_i}{2}>W_i>0)}\right].$$

Hölders inequality with $p = 1$ and $q = \infty$ implies $\sigma E\left[\log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,\sigma)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,\sigma)}\right)\right] \leq -E\left[-\frac{b_i}{2}1_{(b_i<0)}\right] Pr(\frac{b_i}{2} < W_i < 0) - E\left[\frac{b_i}{2}1_{(0<b_i)}\right] Pr(0 < W_i < \frac{b_i}{2})$. □

The next proposition shows that the KL minimizer is the parameter vector that satisfies the subgradient conditions.

Proposition C.1. *Suppose Assumptions 3.2 and 3.5 hold. Then*

$$\inf_{(\alpha,\beta)\in\Theta} E\left[\log\left(\frac{p_0(\mathbf{Y}_i)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,1)}\right)\right] \geq E\left[\log\left(\frac{p_0(\mathbf{Y}_i)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)}\right)\right]$$

with equality if $(\alpha,\beta) = (\alpha_0,\beta_0)$ where (α_0,β_0) are defined in (3.11) and (3.12).

Proof. This follows from the previous lemma and the fact that

$$E\left[\log\left(\frac{p_0(\mathbf{Y}_i)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,1)}\right)\right] = E\left[\log\left(\frac{p_0(\mathbf{Y}_i)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)}\right)\right] + E\left[\log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,1)}\right)\right]$$

□

Now I create an upper bound to approximate $E[I_n(B)^d]$.

Lemma C.3. *Suppose Assumptions 3.3a or 3.3b hold and 3.4 holds. Let $B \subset \Theta \subset \Re^k$. For $\delta > 0$ and $d \in (0,1)$, let $\{A_j : 1 \leq j \leq J(\delta)\}$ be hypercubes of volume $\left(\frac{\delta\frac{1}{k}}{1+c_\Gamma c_y}\right)^k$ required to cover B. Then for $(\alpha^{(j)},\beta^{(j)}) \in A_j$, the following inequality holds*

$$E\left[\left(\int_B \prod_{i=1}^n \frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)} d\Pi(\alpha,\beta)\right)^d\right] \leq \sum_{j=1}^{J(\delta)}\left[E\left[\left(\prod_{i=1}^n \frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_j,\beta_j,1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)}\right)^d\right] e^{nd\delta}\Pi(A_j)^d\right]$$

Proof. For all $(\alpha, \beta) \in A_j$, $|\alpha - \alpha^{(j)}| \leq \frac{\delta^{\frac{1}{k}}}{1+c_\Gamma c_y}$ and $|\beta - \beta^{(j)}| \leq \frac{\delta^{\frac{1}{k}}}{1+c_\Gamma c_y}\mathbf{1}_{k-1}$ compenentwise. Then $|\alpha - \alpha^{(j)}| + |\beta - \beta^{(j)}|'\mathbf{1}_{k-1}c_\Gamma c_y \leq \delta$. Using lemma C.1b

$$\begin{aligned}
\log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha^{(j)},\beta^{(j)},1)}\right) &\leq |\alpha - \alpha^{(j)}| + |\beta - \beta^{(j)}|'|\Gamma_u'||\mathbf{Y}_i| \\
&\leq |\alpha - \alpha^{(j)}| + |\beta - \beta^{(j)}|'\mathbf{1}_{k-1}c_\Gamma c_y \\
&\leq \frac{\delta}{1 + c_\Gamma c_y} \\
&< \delta
\end{aligned}$$

Then $\int_{A_j} \prod_{i=1}^n \frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)} d\Pi(\alpha,\beta) =$

$$\begin{aligned}
&\prod_{i=1}^n \frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha^{(j)},\beta^{(j)},1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)} \int_{A_j} \prod_{i=1}^n \frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha^{(j)},\beta^{(j)},1)} d\Pi(\alpha,\beta) \\
&\leq \prod_{i=1}^n \frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha^{(j)},\beta^{(j)},1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)} e^{n\delta}\Pi(A_j)
\end{aligned}$$

Then $E\left[\left(\int_B \prod_{i=1}^n \frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)} d\Pi(\alpha,\beta)\right)^d\right] \leq$

$$\begin{aligned}
&E\left[\left(\sum_{j=1}^{J(\delta)} \left(\prod_{i=1}^n \frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha^{(j)},\beta^{(j)},1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)} d\Pi(\alpha,\beta)\right) e^{n\delta}\Pi(A_j)\right)^d\right] \\
&\leq \sum_{j=1}^{J(\delta)} E\left[\left(\prod_{i=1}^n \frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha^{(j)},\beta^{(j)},1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)} d\Pi(\alpha,\beta)\right)^d e^{nd\delta}(\Pi(A_j))^d\right]
\end{aligned}$$

the last inequality holds because $(\sum_i x_i)^d \leq \sum_i x_i^d$ for $d \in (0,1)$ and $x_i > 0$. □

Let $U_n^c \subset \Theta$ such that $(\alpha_0, \beta_0) \notin U_n^c$. The next lemma creates an upper bound for the expected value of the likelihood within U_n^c. Break U_n^c into a sequence of half spaces, $\{V_{ln}\}_{l=1}^{L(k)}$,

such that $\bigcup_{l=1}^{L(k)} V_{ln} = U_n^c$, where

$$
\begin{aligned}
V_{1n} &= \{(\alpha,\beta) : \alpha - \alpha_0 \geq \Delta_n, \beta_1 - \beta_{01} \geq 0, ..., \beta_k - \beta_{0k} \geq 0\} \\
V_{2n} &= \{(\alpha,\beta) : \alpha - \alpha_0 \geq 0, \beta_1 - \beta_{01} \geq \Delta_n, ..., \beta_k - \beta_{0k} \geq 0\} \\
&\vdots \\
V_{L(k)n} &= \{(\alpha,\beta) : \alpha - \alpha_0 < 0, \beta_1 - \beta_{01} < 0, ..., \beta_k - \beta_{0k} \leq -\Delta_n\}
\end{aligned}
$$

for some $\Delta_n > 0$. This sequence makes it explicit that at least one component of the vector (α, β) is further than it's corresponding component of (α_0, β_0) by at least an absolute distance Δ_n. How the sequence is indexed exactly is not important. I will focus on V_{1n}, the arguments for the other sets are similar. Define $B_{in} = -E\left[\log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)}\right)\right]$.

Lemma C.4. *Let $G \in \Theta$ be compact. Suppose Assumption 3.4 holds and $(\alpha,\beta) \in G \cap V_{1n}$. Then there exists a $d \in (0,1)$ such that*

$$
E\left[\prod_{i=1}^{n}\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)}\right)^d\right] \leq e^{-d\sum_{i=1}^{n} B_{in}}
$$

Proof. Define $h_d(\alpha,\beta) = \frac{1-E\left[\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)}\right)^d\right]}{d} - E\left[log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)}\right)\right]$. From the proof of Lemma 6.3 in Kleijn and van der Vaart (2006), $\lim_{d\to 0} h_d(\alpha,\beta) = 0$ and $h_d(\alpha,\beta)$ is a decreasing function of d for all (α,β). Note that $h_d(\alpha,\beta)$ is continuous in (α,β). Then by Dini's theorem $h_d(\alpha,\beta)$ converges to $h_d(0,\mathbf{0}_{k-1})$ uniformly in (α,β) as d converges to zero. Define $\delta = \inf_{(\alpha,\beta)\in G} \log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)}\right)$ then there exists a d_0 such that $0 - h_{d_0}(\alpha,\beta) \leq \frac{\delta}{2}$. From lemma

C.2a $E\left[log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)}\right)\right] < 0$. Then

$$\begin{aligned} E\left[\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)}\right)^{d_0}\right] &\leq 1 + d_0 E\left[log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)}\right)\right] + d_0\frac{\delta}{2} \\ &\leq 1 + \frac{d_0}{2} E\left[log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)}\right)\right] \\ &\leq e^{\frac{d_0}{2}E\left[log\left(\frac{f_{\boldsymbol{\tau}}\mathbf{Y}_i|\alpha,\beta,1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)}\right)\right]} \end{aligned}$$

The last inequality holds because $1 + t \leq e^t$ for any $t \in \Re$. □

I would like the thank Karthik Sriram for help with the previous proof. The next lemma is used to show the numerator of the posterior, $I_n(U_n^c)$, converges to zero for sets U_n^c not containing (α_0, β_0).

Lemma C.5. *Suppose Assumptions 3.3a, 3.4 and 3.6 hold. Then there exists a $u_j > 0$ such that for any compact $G_j \subset \Theta$,*

$$\int_{G_j^c \cap V_{jn}} e^{\sum_{i=1}^n \log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)}\right)} d\Pi(\alpha,\beta) \leq e^{-nu_j}$$

for sufficiently large n.

Proof. Let

$$C_0 = \frac{4 \lim_{n\to\infty} \frac{1}{m}\sum_{i=1}^m E[|W_i|]}{(1-\tau)c_p},$$

$\epsilon = \min(\epsilon_Z)$ and $A = kB\epsilon = 2C_0$, where c_p and ϵ_z are from Assumption 3.6. This limit exists by Lemma C.2e. Define

$$G_1 = \{(\alpha,\beta) : (\alpha - \alpha_0, \beta_1 - \beta_{01}, ..., \beta_k - \beta_{0k}) \in [0,A] \times [0,B] \times ... \times [0,B]\}.$$

If $(\alpha,\beta) \in G_1^c \cap W_1$ then $(\alpha-\alpha_0) > A$ or $(\beta-\beta_0)_j > B$ for some j. If $\mathbf{Y}_{\mathbf{u}i}^{\perp} > \epsilon$ then $b_i = (\alpha-\alpha_0) + (\beta-\beta_0)'\mathbf{Y}_{\mathbf{u}i}^{\perp} > 2C_0$. Split the likelihood as

$$\sum_{i=1}^{n} \log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)}\right) =$$
$$\sum_{i=1}^{n} \log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)}\right) 1_{(\mathbf{Y}_{\mathbf{u}ij}^{\perp} > \epsilon_{Zj},\forall j)} + \sum_{i=1}^{n} \log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)}\right) (1 - 1_{(\mathbf{Y}_{\mathbf{u}ij}^{\perp} > \epsilon_{Zj},\forall j)}).$$

Since $\min(W_i^+, b_i) \leq W_i^+ \leq |W_i|$ and using lemma C.1 d,

$$\begin{aligned}\sum_{i=1}^{n} log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)}\right) 1(\mathbf{Y}_{\mathbf{u}ij}^{\perp} > \epsilon_{Zj},\forall j) &= \sum_{i=1}^{n}(-b_i(1-\tau) + min(W_i^+, b_i))1_{(\mathbf{Y}_{\mathbf{u}ij}^{\perp} > \epsilon_{Zj},\forall j)} \\ &\leq \sum_{i=1}^{n}(-2C_0(1-\tau) + |W_i|)1_{(\mathbf{Y}_{\mathbf{u}ij}^{\perp} > \epsilon_{Zj},\forall j)}.\end{aligned}$$

From lemma C.1b and for large enough n then

$$\sum_{i=1}^{n} log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)}\right) 1_{(\mathbf{Y}_{\mathbf{u}ij}^{\perp} > \epsilon_{Zj},\forall j)} \leq \sum_{i=1}^{n} |W_i|(1 - 1_{(\mathbf{Y}_{\mathbf{u}ij}^{\perp} > \epsilon_{Zj},\forall j)}).$$

Then for large enough n

$$\begin{aligned}\sum_{i=1}^{n} log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)}\right) &\leq -nC_0(1-\tau)Pr(\mathbf{Y}_{\mathbf{u}ij}^{\perp} > \epsilon_{Zj},\forall j) + 2n \lim_{n\to\infty} \frac{1}{m}\sum_{i=1}^{m} E[|W_i|] \\ &= -2n \lim_{n\to\infty} \frac{1}{m}\sum_{i=1}^{m} E[|W_i|] \\ &= -\frac{1}{2}nC_0(1-\tau)Pr(\mathbf{Y}_{\mathbf{u}ij}^{\perp} > \epsilon_{Zj},\forall j)\end{aligned}$$

Thus the result holds when $u_i = \frac{1}{2}C_0(1-\tau)Pr(\mathbf{Y}_{\mathbf{u}ij}^{\perp} > \epsilon_{Zj},\forall j)$. □

The next lemma shows the marginal likelihood, $I_n(\Theta)$, goes to infinity at the same rate as the numerator in the previous lemma.

Lemma C.6. *Suppose Assumptions 3.3a and 3.4 holds, then*

$$\int_{\Theta} e^{\sum_{i=1}^{n} \log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)}\right)} d\Pi(\alpha,\beta) \geq e^{-n\epsilon}.$$

Proof. From Lemma C.1e $\log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)}\right) \geq -|b_i| \geq -|\alpha-\alpha_0| - |\beta-\beta_0|'|\Gamma_u||\mathbf{Y}_i|$. Define

$$D_\epsilon = \left\{(\alpha,\beta) : |\alpha-\alpha_0| < \frac{\frac{1}{k}\epsilon}{1+c_\Gamma c_y}, |\beta-\beta_0| < \frac{\frac{1}{k}\epsilon}{1+c_\Gamma c_y}\mathbf{1}_{k-1} \text{ componentwise}\right\}.$$

Then for $(\alpha,\beta) \in V_\epsilon$

$$\begin{aligned}
\log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)}\right) &\geq -|\alpha-\alpha_0| - |\beta-\beta_0|'|\Gamma_u||\mathbf{Y}_i| \\
&\geq -|\alpha-\alpha_0| - |\beta-\beta_0|'\mathbf{1}_{k-1}c_\Gamma c_y \\
&\geq -\frac{\epsilon}{1+c_\Gamma c_y} \\
&> -\epsilon
\end{aligned}$$

Then $\sum_{i=1}^{n} \log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)}\right) \geq -n\epsilon$. If $\Pi(\cdot)$ is proper, then $\Pi(D_\epsilon) \leq 1$.

□

The previous two lemmas imply the posterior is converging to zero in a restricted parameter space.

Lemma C.7. *Suppose Assumptions 3.4, and 3.6 hold. Then for each $l \in \{1,2,...,L(k)\}$, there exists a compact G_l such that*

$$\lim_{n\to\infty} \Pi(V_{ln} \cap G_l^c | \mathbf{Y}_1,...,\mathbf{Y}_n) = 0.$$

Proof. Let ϵ from Lemma C.6 equal $\frac{u_i}{4}$ from Lemma C.5. Then

$$\int_\Theta e^{\sum_{i=1}^n log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)}\right)} d\Pi(\alpha,\beta) \geq \int_{D_\epsilon} e^{\sum_{i=1}^n log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)}\right)} d\Pi(\alpha,\beta)$$
$$\geq e^{-n\epsilon} d\Pi(D_\epsilon)$$

Then $\lim_{n\to\infty} \int_\Theta e^{\sum_{i=1}^n log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)}\right)} d\Pi(\alpha,\beta) e^{nu_j/2} = \infty$ and

$\lim_{n\to\infty} \int_{V_{jn}\cap G_j^c} e^{\sum_{i=1}^n log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)}\right)} d\Pi(\alpha,\beta) e^{nu_j/2} = 0.$ □

The next proposition bounds the expected value of the numerator, $E[I_n(V_{1n}\cap G)^d]$, and the denominator, $I_n(\Theta)$, of the posterior. Define $B_{in} = -E\left[\log\left(\frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)}\right)\right]$.

Lemma C.8. *Suppose Assumptions 3.3a and 3.4 hold. Define* $D_{\delta_n} = \left\{(\alpha,\beta) : |\alpha-\alpha_0| < \frac{\frac{1}{k}\delta_n}{1+c_\Gamma c_y}, |\beta-\beta_0| < \frac{\frac{1}{k}\delta_n}{1+c_\Gamma c_y}\mathbf{1}_{k-1} \text{ componentwise}\right\}$. *Then for* $(\alpha,\beta)\in D_{\delta_n}$

1. There exists a $\delta_n \in (0,1)$ *and fixed* $R>0$ *such that*

$$E\left[\left(\int_{V_{1n}\cap G}\prod_{i=1}^n \frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)} d\Pi(\alpha,\beta)\right)^d\right] \leq e^{d\sum_{i=1}^n B_{in}} e^{nd\delta_n} R^2/\delta_n^2$$

2.

$$\int_\Theta \prod_{i=1}^n \frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)} d\Pi(\alpha,\beta) \geq e^{-n\delta_n}\Pi(D_{\delta_n})$$

Proof. From Lemma C.3 and C.4 $E\left[\left(\int_{W_{1n}\cap G}\prod_{i=1}^n \frac{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha,\beta,1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)} d\Pi(\alpha,\beta)\right)^d\right]$

$$
\begin{aligned}
&\leq \sum_{j=1}^{J(\delta_n)} \left[E\left[\left(\prod_{i=1}^{n} \frac{f_{u,\tau}(\mathbf{Y}_i|\alpha_j,\beta_j,1)}{f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)} \right)^d \right] e^{nd\delta_n} \Pi(A_j)^d \right] \\
&\leq \sum_{j=1}^{J(\delta_n)} \left[e^{-d\sum_{i=1}^{n} B_{in}} e^{nd\delta_n} \Pi(A_j)^d \right] \\
&\leq e^{-d\sum_{i=1}^{n} B_{in}} e^{nd\delta_n} J(\delta_n)
\end{aligned}
$$

Since G is compact, R can be chosen large enough so that $J(\delta_n) \leq R^2/\delta_n^2$. 2) is from Lemma C.7. □

The proof of Theorem 3.1 is below.

Proof. Suppose Π is proper. Lemma C.5 shows we can focus on the case $W_{1n} \cap G$. Set $\Delta_n = \Delta$ and $\delta_n = \delta$. Then from Lemma C.8, there exists a $d \in (0,1)$ such that for sufficiently large n

$$
\begin{aligned}
E\left[(\Pi(V_{1n} \cap G|\mathbf{Y}_1,...,\mathbf{Y}_n))^d \right] &\leq \frac{R^2}{\delta^2 (\Pi(V_\delta))^d} e^{-d\sum_{i=1}^{n} B_{in}} e^{2nd\delta} \\
&\leq \frac{R^2}{\delta^2 (\Pi(V_\delta))^d} e^{-\frac{1}{2}dn \lim_{m\to\infty} \frac{1}{m}\sum_{i=1}^{m} B_{im}} e^{2nd\delta}
\end{aligned}
$$

Chose $\delta = \frac{1}{8} \lim_{m\to\infty} \frac{1}{m} \sum_{i=1}^{m} B_{im}$ and note that $C' = \frac{R^2}{\delta^2 (\Pi(V_\delta))^d}$ is a fixed constant. Then $E\left[(\Pi(V_{1n} \cap G|\mathbf{Y}_1,...,\mathbf{Y}_n))^d \right] \leq C' e^{-nd\delta/4}$. Since $\lim_{n\to\infty} \sum_{n=1}^{\infty} C' e^{-nd\delta/4} < \infty$ then the Markov inequality and Borel Cantelli imply posterior consistency a.s..

Now suppose the prior is improper but admits a proper posterior. Consider the posterior from the first observation $\Pi(\cdot|\mathbf{Y}_1)$. Under Assumption 3.3b, $\Pi(\cdot|\mathbf{Y}_1)$ is proper. Assumption 3.5 ensures that $f_{\boldsymbol{\tau}}(\mathbf{Y}_i|\alpha_0,\beta_0,1)$ dominates p_0. Thus the formal posterior exists on a set of $\mathbf{P}$ measure 1. Further, $\Pi(U|\mathbf{Y}_1) > 0$ for some open U containing (α_0,β_0). Thus $\Pi(\cdot|\mathbf{Y}_1)$ can

be used as a proper prior on the likelihood containing $\mathbf{Y}_2, ..., \mathbf{Y}_n$ which produces a posterior equivalent to the original $\Pi(\cdot|\mathbf{Y}_1, ..., \mathbf{Y}_n)$ and thus the same argument above using a proper prior can be applied to the posterior $\Pi(\cdot|\mathbf{Y}_2, ..., \mathbf{Y}_n)$ using $\Pi(\cdot|\mathbf{Y}_1)$ as a proper prior. □

I would like to thank Karthik Sriram and R.V. Ramamoorthi for help with the improper prior case.

C.2 Non-zero centered prior: second approach

The second approach is to investigate the implicit prior in the untransformed response space of Y_2 against Y_1, $\mathbf{X}$ and an intercept. Denote $\mathbf{\Gamma_u} = [u_1^\perp, u_2^\perp]'$. Note that $\mathbf{Y_{u}}_i = \beta_{\boldsymbol{\tau}\mathbf{y}}\mathbf{Y}_{\mathbf{u}i}^\perp + \beta'_{\boldsymbol{\tau}\mathbf{x}}\mathbf{X}_i + \alpha_{\boldsymbol{\tau}}$ can be rewritten as

$$\begin{aligned} Y_{2i} &= \frac{1}{u_2 - \beta_{\boldsymbol{\tau}\mathbf{y}} u_2^\perp}\left((\beta_{\boldsymbol{\tau}\mathbf{y}} u_1^\perp - u_1)Y_{1i} + \beta'_{\boldsymbol{\tau}\mathbf{x}}\mathbf{X}_i + \alpha_{\boldsymbol{\tau}}\right) \\ &= \phi_{\boldsymbol{\tau}y}Y_{1i} + \phi'_{\boldsymbol{\tau}\mathbf{x}}\mathbf{X}_i + \phi_{\boldsymbol{\tau}1} \end{aligned}$$

Since the equation is in slope-intercept form, the interpretation of $\phi_{\boldsymbol{\tau}}$ is fairly straight forward. It can be verified that $\phi_{\boldsymbol{\tau}y} = \phi_{\boldsymbol{\tau}y}(\beta_{\boldsymbol{\tau}\mathbf{y}}) = \frac{\beta_{\boldsymbol{\tau}\mathbf{y}}u_1^\perp - u_1}{u_2 - \beta_{\boldsymbol{\tau}\mathbf{y}}u_2^\perp} = \frac{1}{u_1(u_2^\perp \beta_{\boldsymbol{\tau}\mathbf{y}} - u_2)} + \frac{u_2}{u_1}$ for $\beta_{\boldsymbol{\tau}y} \neq \frac{u_2}{u_2^\perp}$ and $u_1 \neq 0$. Suppose prior $\theta_{\boldsymbol{\tau}} = [\beta_{\boldsymbol{\tau}\mathbf{y}}, \beta'_{\boldsymbol{\tau}\mathbf{x}}, \alpha_{\boldsymbol{\tau}}]' \sim F_{\theta_{\boldsymbol{\tau}}}(\theta_{\boldsymbol{\tau}})$ with support $\Theta_{\boldsymbol{\tau}}$. If $F_{\beta\mathbf{y}}$ is a continuous distribution, the density of $\phi_{\boldsymbol{\tau}}$ is

$$f_{\phi_{\boldsymbol{\tau}y}} = f_{\beta_{\boldsymbol{\tau}\mathbf{y}}}(\phi_{\boldsymbol{\tau}y}^{-1}(\beta_{\boldsymbol{\tau}\mathbf{y}}))\left|\frac{d}{d\beta_{\boldsymbol{\tau}\mathbf{y}}}\phi_{\boldsymbol{\tau}y}^{-1}(\beta_{\boldsymbol{\tau}\mathbf{y}})\right| = f_{\beta_{\boldsymbol{\tau}\mathbf{y}}}\left(\frac{1}{u_2^\perp}\left(\frac{1}{u_1\phi_{\boldsymbol{\tau}y} - u_2} + u_2\right)\right)\left|\frac{u_1}{u_2^\perp(u_1\phi_{\boldsymbol{\tau}y} - u_2)^2}\right|$$

with support not containing $\left\{-\frac{u_1^\perp}{u_2^\perp}\right\}$, for $u_2^\perp \neq 0$.

If $\beta_{\boldsymbol{\tau}\mathbf{y}} \sim N(\underline{\mu}_{\boldsymbol{\tau}y}, \underline{\sigma}^2_{\boldsymbol{\tau}y})$, then the density of $\phi_{\boldsymbol{\tau}y}$ is a shifted reciprocal Gaussian with density

$$f_{\phi_{\boldsymbol{\tau}y}}(\phi|\underline{a}, \underline{b}^2) = \frac{1}{\sqrt{2\pi\underline{b}^2_{\boldsymbol{\tau}}(\phi - u_2/u_2^{\perp})^2}} exp\left(-\frac{1}{2\underline{b}^2_{\tau}}\left(\frac{1}{\phi - u_2/u_2^{\perp}} - \underline{a}\right)^2\right).$$

The parameters are $\underline{a} = \underline{\mu}_{\boldsymbol{\tau}} u_1 u_2^{\perp} - u_1 u_2$ and $\underline{b} = u_1 u_2^{\perp} \underline{\sigma}_{\boldsymbol{\tau}}$. The moments of $\phi_{\boldsymbol{\tau}y}$ do not exist (Robert, 1991). The density is bimodal with modes at

$$m_1 = \frac{-\underline{a} + \sqrt{\underline{a}^2 + 8\underline{b}^2}}{4\underline{b}^2} + \frac{u_2}{u_2^{\perp}} \text{ and } m_2 = \frac{-\underline{a} - \sqrt{\underline{a}^2 + 8\underline{b}^2}}{4\underline{b}^2} + \frac{u_2}{u_2^{\perp}}.$$

Since moments do not exist, calibration can be tricky and has to rely on the modes and their relative heights

$$\frac{f_{\phi_{\boldsymbol{\tau}y}}(m_1|\underline{a}, \underline{b}^2)}{f_{\phi_{\boldsymbol{\tau}y}}(m_2|\underline{a}, \underline{b}^2)} = \frac{\underline{a}^2 + \underline{a}\sqrt{\underline{a}^2 + 8\underline{b}^2} + 4\underline{b}^2}{\underline{a}^2 - \underline{a}\sqrt{\underline{a}^2 + 8\underline{b}^2} + 4\underline{b}^2} exp\left(\frac{\underline{a}\sqrt{\underline{a}^2 + 8\underline{b}^2}}{\underline{b}^4}\right)$$

A few plots of the Reciprocal Gaussian are shown in figure C.1.

The distribution of $\phi_{\boldsymbol{\tau}\mathbf{x}}$ and $\phi_{\boldsymbol{\tau}1}$ are ratio normals. I will discuss the implied prior on $\phi_{\boldsymbol{\tau}1}$. The distribution of $\phi_{\boldsymbol{\tau}\mathbf{x}}$ will follow by analogy. The implied intercept $\phi_{\boldsymbol{\tau}1} = \frac{\alpha_{\boldsymbol{\tau}}}{u_2 - \beta_{\boldsymbol{\tau}\mathbf{y}} u_2^{\perp}}$ is a ratio of normals distribution. The ratio of normals distributions can always be expressed as a location scale shift of $R = \frac{Z_1 + a}{Z_2 + b}$ where $Z_i \overset{iid}{\sim} N(0, 1)$ for $i \in \{1, 2\}$. That is there exist constants c and d such that $\phi_{\boldsymbol{\tau}1} = cR + d$ (Marsaglia, 1965; Hinkley, 1969, 1970).[1] The

[1] Let $W_i \sim N(\theta_i, \sigma_i^2)$ for $i \in \{1, 2\}$ with $corr(W_1, W_2) = \rho$. Then $\frac{W_1}{W_2} = \frac{\sigma_1}{\sigma_2}\sqrt{1 - \rho^2}\left(\frac{\frac{\theta_1}{\sigma_1} + Z_1}{\frac{\theta_2}{\sigma_2} + Z_2} + \frac{\rho}{\sqrt{1-\rho^2}}\right)$ where $Z_i \sim N(0, 1)$ for $i \in \{1, 2\}$ with $corr(Z_1, Z_2) = 0$. Thus $a = \frac{\theta_1}{\sigma_1}$, $b = \frac{\theta_2}{\sigma_2}$, $c = \frac{\sigma_1}{\sigma_2}\sqrt{1 - \rho^2}$ and $d = c\frac{\rho}{\sqrt{1-\rho^2}}$ where $\theta_1 = \underline{a}_{\boldsymbol{\tau}1}$, $\theta_2 = u_2 - \underline{a}_{\boldsymbol{\tau}y} u_2^{\perp}$, $\sigma_1 = \underline{b}_{\boldsymbol{\tau}1}$ and $\sigma_2 = \underline{b}_{\boldsymbol{\tau}y} u_2^{\perp}$.

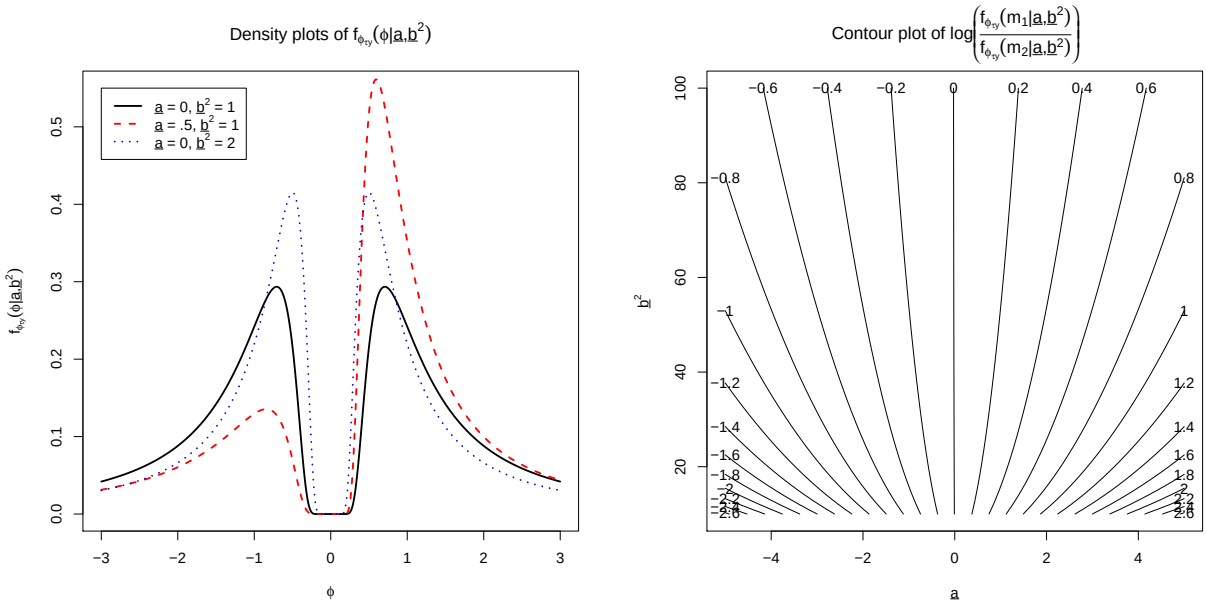

Figure C.1: (left) Density of $f_{\phi_{\boldsymbol{\tau} y}}(\phi|\underline{a},\underline{b}^2)$ for hyper parameters $\underline{a}=0$, $\underline{b}^2=1$ (solid black), $\underline{a}=0.5$, $\underline{b}^2=1$ (dash red), $\underline{a}=0$, $\underline{b}^2=2$ (dotted blue). (right) A contour plot showing the logged relative heights of the modes at m_1 over m_2 over the grid $(\underline{a},\underline{b}^2)\in[-5,5]\times[10,100]$.

density of $\phi_{\boldsymbol{\tau}1}$ is

$$f_{\phi_{\boldsymbol{\tau}1}}(\phi|\underline{a},\underline{b}) = \frac{e^{-\frac{1}{2}(\underline{a}^2+\underline{b}^2)}}{\pi(1+\phi^2)}\left[1+ce^{\frac{1}{2}c^2}\int_0^c e^{-\frac{1}{2}t^2}dt\right], \text{ where } c=\frac{\underline{b}+\underline{a}\phi}{\sqrt{1+\phi^2}}.$$

When $\underline{a}=\underline{b}=0$, then the distribution reduces down to the standard cauchy distribution. The distribution, like the reciprocal normal distribution, has no moments and can be bimodal. Unlike the reciprocal normal, there does not exist a closed form solution for the exact location of the modes. Focusing on the positive quadrant of $(\underline{a},\underline{b})$, if $a\leq 1$ then the distribution is unimodal and if $a>2.256058904$ then the distribution is bimodal (discussion of the other quadrants is relegated to the appendix). There is a curve that separates the two regions as shown in the bottom right of figure C.2.[2] If the distribution is bimodal, one mode will be to the left of $-\underline{b}/\underline{a}$ and the other to the right. The left mode tends to be much lower than the right for positive $(\underline{a},\underline{b})$. Unlike the reciprocal gaussian closed form solutions for the modes do not exist. However, the distribution is approximately elliptical with 'central

[2]The curve is approximately $\underline{b}=\frac{18.621-63.411\underline{a}^2-54.668\underline{a}^3+17.716\underline{a}^4-2.2986\underline{a}^5}{2.256058904-\underline{a}}$ for $1\leq 2.256....$

tendency' $\mu = \frac{\underline{a}}{1.01\underline{b}-0.2713}$ and 'squared dispersion' $\sigma^2 = \frac{\underline{a}^2+1}{\underline{b}^2+0.108\underline{b}-3.795} - \mu^2$ when $\underline{a} < 2.256$ and $4 < \underline{b}$ (Marsaglia, 2006).

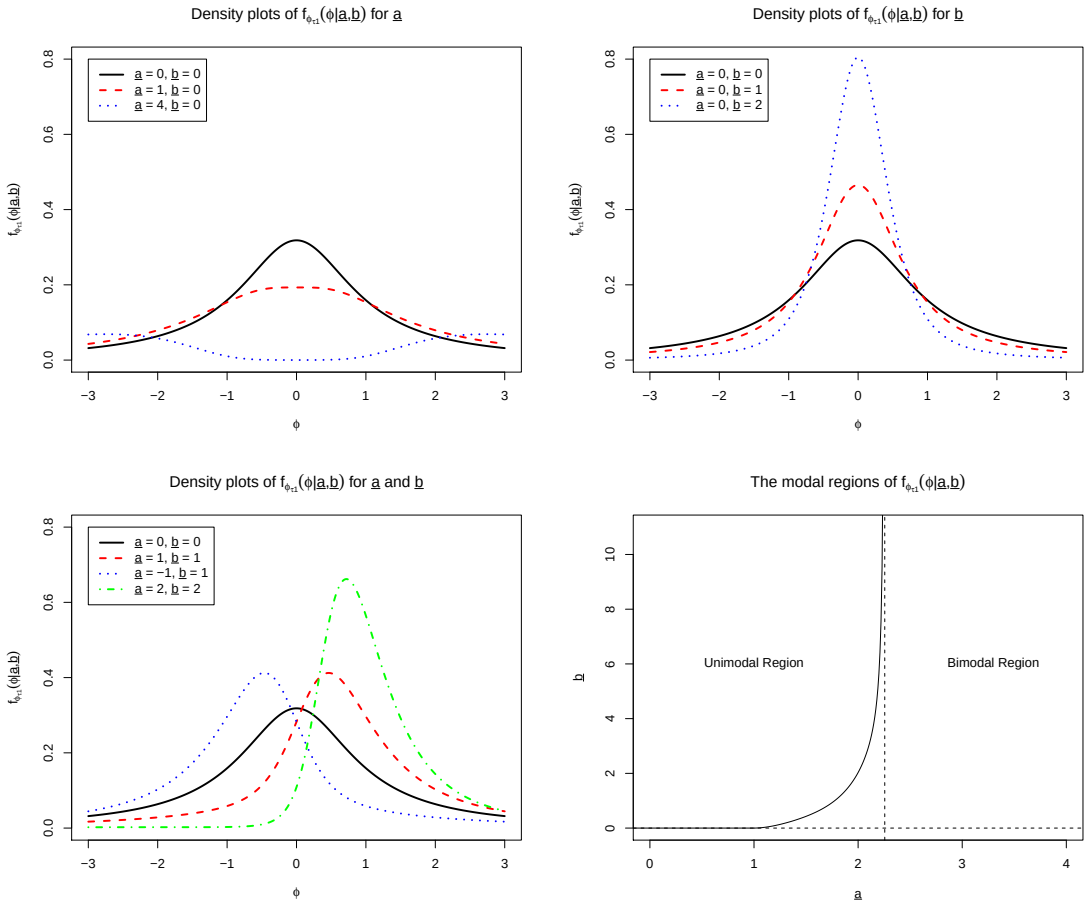

Figure C.2: The top two plots and the bottom left plot show the density of the ratio normal distribution with parameters $(\underline{a}, \underline{b})$. The top left plot shows the density for different values of $\underline{a}$ with $\underline{b}$ fixed at zero. The parameters $(\underline{a}, \underline{b}) = (1, 0)$ and $(4, 0)$ result in the same density as $(\underline{a}, \underline{b}) = (-1, 0)$ and $(-4, 0)$. The top right plot shows the density for different values of $\underline{b}$ with $\underline{a}$ fixed at zero. The parameters $(\underline{a}, \underline{b}) = (0, 1)$ and $(0, 2)$ result in the same density as $(\underline{a}, \underline{b}) = (0, -1)$ and $(0, -2)$. The bottom left plot shows the density for different values of $\underline{a}$ and $\underline{b}$. The parameters $(\underline{a}, \underline{b}) = (1, 1)$, $(-1, 1)$and $(2, 2)$ result in the same density as $(\underline{a}, \underline{b}) = (-1, -1)$, $(1, -1)$and $(-2, -2)$. The bottom right graph shows the regions of the positive quadrant of the parameter space where the density is either bimodal or unimodal.

Appendix D

Chapter 4

D.1 Lemmas and proofs

This lemma provides distributions for C_X, C_Y, and $C_X + C_Y$.

Lemma D.1. *Suppose Assumption 4.1 then* $C_X + C_Y|X, \mathbf{p} \sim$ *Poisson-Binomial*$(\mathbf{p})$ *where* $\mathbf{p} = (p_1, p_2, ..., p_9)$. *This result still holds for Assumption 1.1b as well.*

The proof of Theorem 4.1 is below.

Proof. From Lemma D.1, the distribution of $C_X + C_Y|X, \mathbf{p}$ is not a function of X. Thus $Pr(C_X + C_Y|X, \mathbf{p}) = Pr(C_X + C_Y|\mathbf{p})$ and independence is maintained after integrating out $\mathbf{p}$. □

The proof of Lemma 4.1 is below.

Proof. Without loss of generality let justices $1, 2, ..., X$ be the justices affirming. If $X = 0$ then $1, 2, ..., 9$ are voting to reverse. Since $c_i|p_i \sim Bernoulli(p_i)$ then $C_X|X, p_1, p_2, ...p_X =$

$\sum_{i=1}^{X} c_i|X, p_1, p_2, ...p_X$. Then the probability of $C_X = k|X, p_1, p_2, ..., p_X$ is

$$Pr(C_X = k|X, p_1, p_2, ..., p_X) = \sum_{A \in \mathcal{F}_k} \prod_{i \in A} p_i \prod_{j \in A^c} (1 - p_j)$$

where $\mathcal{F}_k$ is the set of all subsets of size k from $\{1, 2, ..., X\}$ (Wang, 1993). It follows that $Pr(C_X = k|X) = \int Pr(C_X = k|X, p_1, p_2, ..., p_X) dF_{p_1,p_2,...,p_X} = \int \dots \int Pr(C_X = k|X, p_1, p_2, ..., p_X) dF_{p_1} dF_{p_2}$. Notice for any given $A \in \mathcal{F}_k$, $\prod_{i \in A} p_i$ and $\prod_{j \in A^c}(1 - p_j)$ have no common p_l for any l. It follows that $Pr(C_X = k|X) = \sum_{A \in \mathcal{F}_k} \prod_{i \in A} \mu_p \prod_{j \in A^c}(1 - \mu_p) = \binom{X}{k} \mu_p^k (1 - \mu_p)^{X-k}$. Thus $C_X|X \sim Binomial(X, \mu_p)$. Likewise, $C_Y|X \sim Binomial(9 - X, \mu_p)$ and $C_X + C_Y \sim Binomial(9, \mu_p)$. Since $C_X|X \perp C_Y|X$ it follows that the probability mass function of $C_X|X, C_X+C_Y$ is $\frac{\binom{X}{C_x}\binom{9-X}{C_Y}}{\binom{9}{C_X+C_Y}}$ which means $C_X|X, C_X+C_Y \sim Hypergeometric(9, X, C_X+C_Y)$. It follows that $Pr(C_X = 0|X, C_X + C_Y = 1) = 1 - \frac{X}{9}$. □

The proof of Lemma 4.2 is below.

Proof. The conditional expectation is

$$\begin{aligned} E[X|X^\star] &= X^\star Pr(X = X^\star|X^\star) + (X^\star + 1)\,(1 - Pr(X = X^\star|X^\star)) \\ &= X^\star + 1 - Pr(X = X^\star|X^\star) \end{aligned}$$

Since conditioning on $X^\star$ truncates X then $Pr(X = X^\star|X^\star) = \frac{q_{X^\star}}{q_{X^\star} + q_{X^\star+1}}$. The result follows. □

The proof of Theorem 4.2 is below.

Proof. $X^\star + Y^\star = 8$ means there was a recusal. It follows from assumption 4.3 that $C_X + C_Y \geq 1$. Since $C_X + C_Y \in \{0, 1\}$ then assumption 4.2 is true and $C_X + C_Y = 1$. By assumption 4.2, $X|(R = 1) = X|(C_X + C_Y = 1)$. By Theorem 4.1, $X|C_X + C_Y = 1 \sim X$. Thus

$X|R = 1 \sim X$. □

D.2 Supreme court process

The United States Supreme Court is the highest level of judicial authority in the United States. It is composed of nine justices who rule on the cases presented before the court. An odd number of justices was chosen to prevent tie votes. Justices are chosen by the President and are confirmed by the Senate. Justices hold their seat at the court until death or they decide to step down.

The Supreme Court holds original jurisdiction over disputes between states, ambassadors or other high ranking ministers.[1] This is a minority of the cases that the court hears (recently about 1 or 2 a year). Most cases are appellate cases. These are cases from lower courts where one of the parties involved are unsatisfied with the ruling or process and submit a Writ of Certiorari to a higher court asking to review the case. Since the Supreme Court is the highest court, there are no appeals from their rulings. The Supreme Court is not required to review every case that submits a Writ of Certiorari. The court accepts less than 100 of the 7000 requests submitted to them annually. A case is accepted if at least 4 of the 9 justices vote to hear the case.[2] If a case is not accepted, the ruling of the lower court is accepted as the final ruling.[3]

The parties then gather in front of the court and present oral arguments. The oral arguments

[1] Original jurisdiction means the first court to hear the matter. Since they are the highest authority, they are the only court to hear matters of original jurisdiction.

[2] 7000 requests is too many for the judges to each individually read and vote on in an efficient manner. So the Writs of Certiorari are divided among the law clerks (there are 3 or 4 clerks per justice) who write a summary and provide a recommendation to whether the Supreme Court should hear the case. Then once a week the justices get together and vote on cases based on the compiled summaries and recommendations. Most justices pool together the certs to share the workload among the clerks. Some justices do not participate in this pool.

[3] If a case is accepted then the case is placed on the docket and the petitioner must write a brief describing the issue that the court has agreed to review. Then the respondent submits a reply to the brief. Then both parties submit replies to the submitted briefs.

are mainly used so justices can ask the parties questions about the submitted briefs. At the end of the week after oral arguments, the justices meet and discuss the case. After discussions, a vote is held.[4] Then the most senior justice in the majority assigns a justice to write the opinion of the court. If a justice agrees with the decision of the court but not the opinion they can write a concurring opinion. If a justice disagrees with the decision of the court they can write a dissenting opinion. Then the decision of the court and opinions are presented. The decision of the court is not finalized until the majority opinion is signed and presented. There are some rare cases where a justice changes their vote after reading dissenting opinions (e.g. Planned Parenthood v. Casey, 505 U.S. 833).[5]

[4] This vote is sometimes called the conference vote

[5] For a more detailed explanation of the entire process, see http://www.uscourts.gov/about-federal-courts/educational-resources/about-educational-outreach/activity-resources/ supreme-1