DISSERTATION

NONPARAMETRIC TESTS FOR INFORMATIVE SELECTION AND SMALL AREA ESTIMATION FOR RECONCILING SURVEY ESTIMATES

Submitted by

Teng Liu

Department of Statistics

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Summer 2020

Doctoral Committee:

Advisor: F. Jay Breidt

Haonan Wang
Donald J. Estep
Paul F. Doherty, Jr.

ABSTRACT

NONPARAMETRIC TESTS FOR INFORMATIVE SELECTION AND SMALL AREA ESTIMATION FOR RECONCILING SURVEY ESTIMATES

Two topics in the analysis of complex survey data are addressed: testing for informative selection and addressing temporal discontinuities due to survey redesign.

Informative selection, in which the distribution of response variables given that they are sampled is different from their distribution in the population, is pervasive in modern complex surveys. Failing to take such informativeness into account could produce severe inferential errors, such as biased parameter estimators, wrong coverage rates of confidence intervals, incorrect test statistics, and erroneous conclusions. While several parametric procedures exist to test for informative selection in the survey design, it is often hard to check the parametric assumptions on which those procedures are based. We propose two classes of nonparametric tests for informative selection, each motivated by a nonparametric test for two independent samples. The first nonparametric class generalizes classic two-sample tests that compare empirical cumulative distribution functions, including Kolmogorov–Smirnov and Cramér–von Mises, by comparing weighted and unweighted empirical cumulative distribution functions. The second nonparametric class adapts two-sample tests that compare distributions based on the maximum mean discrepancy to the setting of weighted and unweighted distributions. The asymptotic distributions of both test statistics are established under the null hypothesis of noninformative selection. Simulation results demonstrate the usefulness of the asymptotic approximations, and show that our tests have competitive power with parametric tests in a correctly specified parametric setting while achieving greater power in misspecified scenarios.

Many surveys face the problem of comparing estimates obtained with different methodology, including differences in frames, measurement instruments, and modes of delivery. Differences may exist within the same survey; for example, multi-mode surveys are increasingly common. Further, it is inevitable that surveys need to be redesigned from time to time. Major redesign of survey processes could affect survey estimates systematically, and it is important to quantify and adjust for such discontinuities between the designs to ensure comparability of estimates over time. We propose a small area estimation approach to reconcile two sets of survey estimates, and apply it to two surveys in the Marine Recreational Information Program (MRIP). We develop a log-normal model for the estimates from the two surveys, accounting for temporal dynamics through regression on population size and state-by-wave seasonal factors, and accounting in part for changing coverage properties through regression on wireless telephone penetration. Using the estimated design variances, we develop a regression model that is analytically consistent with the log-normal mean model. We use the modeled design variances in a Fay-Herriot small area estimation procedure to obtain empirical best linear unbiased predictors of the reconciled effort estimates for all states and waves, and provide an asymptotically valid mean square error approximation.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my gratitude to my advisor, Dr. F. Jay Breidt, for his guidance and support in my research and study. I am fortunate enough to work with a nice human being who is easy to get along with and cares about his students. Without his help, I would not be able to finish my dissertation the way it is.

Secondly, I would like to thank my committee members: Dr. Haonan Wang, Dr. Don Estep, Dr. Paul Doherty, and my ex-committee member Dr. Jean Opsomer, for serving in my committee and providing valuable advice.

Chapter 4 of this dissertation was supported by NOAA Fisheries. I want to thank scientists Rob Andrews and John Foster for assistance with the problem formulation and data compilation. NOAA affiliate Ryan Kitts-Jensen and Westat statistician Mike Brick provided useful discussion.

Last but not least, I would like to thank Dr. Mary Meyer. Her two emails brought me from China to United States and later to the Ph.D. program. I deeply appreciate the opportunities she provided to me.

DEDICATION

To my son Leonardo

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

1.1 Motivation

1.1.1 Sources of Errors in Surveys

Surveys are used to collect information about a part of a finite population (of people, households, land segments, account records, image pixels, or any other identifiable elements) in order to make inferences about the whole population. Surveys are complex because populations are complex: heterogeneous, hard to identify, hard to access, hard to measure, etc. Several errors could occur during the process of complex surveys. Those errors can be divided into sampling error versus nonsampling error. Sampling error is the error made by drawing inference about the population from only a sample. Nonsampling errors are all other errors, including coverage error, nonresponse error, measurement error and processing error. See section 1.7 of Särndal et al. (1992).

Surveys start with a definition of the population of interest and collection of one or more sampling frames that allow identification and access to population elements: for example, maps, address files, telephone directories, etc. Population elements are sampled from the frame, often using complex procedures that reflect both practical and theoretical considerations. Next, the survey attempts to collect data from the sampled elements. Surveys of human populations use one or more modes of data collection: mail, internet, telephone, or face-to-face interviewing, for example. Not all sampled elements respond, and not all responding elements give accurate data. Finally, the collected data are adjusted statistically, to account for the various sources of error, and used to make inferences about the finite population or the “superpopulation” model assumed to have generated the finite population.

Coverage error occurs when the sampling frame fails to match the target population. Of particular concern is undercoverage, in which some elements are in the population of interest but not in the frame, and cannot be sampled. For example, older telephone surveys were based on landlines in households. However, with the expansion of mobile-only users, landline surveys gradually lose coverage of the targets (Curtin et al. (2005)).

Nonresponse error, in which a sampled element fails to provide some or all data elements, is a challenge in surveys. In certain surveys of human populations, face-to-face interviewing might increase the response rate while in other cases, respondents may refuse to respond or give false information for privacy. If the responses are not missing at random, it could severely bias the inference unless appropriate adjustments are made. See section 14.10 of Särndal et al. (1992).

Measurement error refers to differences between the true data values and the responses recorded in surveys, especially systematic bias. Different modes have to adjust the wording and length of survey questionnaires accordingly, which could lead to bias. For example, a telephone survey may need to break a long question into pieces for the convenience of respondents, and the overall survey cannot take too long. Mail survey respondents have the flexibility to answer longer questions and longer questionnaires. These differences in the format and layout could lead to systematic differences between the responses to the same question presented via different modes. See section 14.8 of Särndal et al. (1992).

Processing error occurs in analyzing the sample, which includes all kinds of errors, such as data entry error, coding error, imputation error, wrong formulas, etc. See section 1.7.iii of Särndal et al. (1992).

This dissertation focuses on two primary problems: informative selection, which may come from some combination of coverage error, sampling error, and nonresponse error; and reconciliation of survey estimates from surveys with different kinds of nonsampling errors.

1.1.2 Informative Selection

By design, complex surveys often involve unequal probabilities of selection, especially to achieve efficiencies, ensure representation of subpopulations, etc. In addition, actual probabilities of selection can differ from designed probabilities of selection due to coverage errors and nonresponse. For simplicity, we will consider the case of known (designed) selection probabilities, though in practice we usually work with weights that are, at least approximately, estimated versions of inverse selection probabilities, reflecting both design and non-design features.

One direct result of unequal probability sampling is that the model holding for sample data can be quite different from the model for the population. Suppose we have a finite population consisting of N elements denoted by a set of indices, $U = \{1, \ldots, N\}$. We denote the response variable of interest by y_i for $i \in U$ and assume that these values in the finite population are generated as independent and identically distributed (iid) realizations from the probability density function (pdf) $f(y_i)$, known as the superpopulation model.

Let $s \subset U$ be the sample of selected elements, and define the sample membership indicator $\xi_i = 1$ if $i \in s$ and $\xi_i = 0$, otherwise. We denote the (unconditional) inclusion probability by $\pi_i = \mathbb{P}(i \in s) = \mathbb{P}(\xi_i = 1)$. The sample pdf is defined as the conditional density of the response, given that it was selected:

$$f_s(y_i) = f(y_i \mid \xi_i = 1) = \frac{\mathbb{P}(\xi_i = 1 \mid y_i) f(y_i)}{\mathbb{P}(\xi_i = 1)}. \tag{1.1.1}$$

It follows from (1.1.1) that population and sample pdfs can be different unless $\mathbb{P}(\xi_i = 1 \mid y_i) = \mathbb{P}(\xi_i = 1)$; that is, the inclusion probability is independent of the variable of interest. If the inclusion probability depends on the variable of interest, we call the sampling design an informative design.

If the design is noninformative, we can ignore the randomness in the design and generalize inference directly from the sample model to the population model. Applying standard analysis to the sample yields a valid inference for the population. Otherwise, we lose efficiency incorporating the design into our analysis.

On the other hand, if the design is informative, the analysis has to be adjusted to get appropriate results. Failure to account for informativeness can lead to biased and inconsistent parameter estimators, poor coverage of confidence intervals, false predictions and other fundamental inferential errors. Several methods have been established to account for informativeness in analysis, such as the design-based Horvitz Thompson (HT) estimation in Horvitz & Thompson (1952). This includes several cases: the parameter is a linear function of random variable; the nonlinear parameter can be explicitly written as a function of random variable; the parameter is solution to a population-level estimating equation; the parameter is the maximum likelihood estimator (MLE) of the population-level model (see Binder (1983)). There are also likelihood-based approaches, e.g., pseudo-likelihood Krieger & Pfeffermann (1992), sample likelihood Patil & Rao (1978) and full likelihood Skinner (1994). Pfeffermann & Sverchkov (1999, 2003) develop sample likelihood approaches under various models.

Thus, testing for informative selection is crucial in analyzing complex surveys. A number of parametric tests exist for testing informative selection. Pfeffermann & Sverchkov (1999) focus on whether the moments of the population model residuals are equal to the moments of the sample model residuals. Another important class of tests is based on assessing the significance of the difference between weighted and unweighted estimators of model parameters. DuMouchel & Duncan (1983) construct a test to compare the weighted and unweighted parameter estimators in a linear model. Fuller (1984) considers the case of cluster samples within strata, and gives an approximate $\mathcal{F}$-test. Pfeffermann (1993) extends the comparison of weighted and unweighted esitmators to general likelihood-based problems with explicit estimators, and provides a Wald-type test statistic. Pfeffermann & Sverchkov (2003) extend that test to estimators that are defined as the solutions to estimating equations in generalized linear models. Herndon (2014) develops the Wald-type test by a parametric bootstrap approach which avoids estimation of covariance matrices.

The work presented in this dissertation extends these weighted/unweighted comparisons from the parametric setting to the nonparametric setting, via distribution functions in Chapter 2 and via maximum mean discrepancy in Chapter 3.

1.1.3 Reconciliation of Estimates from Different Surveys

Different surveys of the same population may use different methods, including frames, sampling designs, nonresponse follow-up, measurement instruments and modes of delivery. This is true of independent surveys, conducted by different investigators, as well as of repeated surveys over time, conducted by the same investigator with periodic methodological changes. Ideally, these differences in methodology would have no effect on inferences about the population. However, these differences in methodology lead to differences in coverage, sampling, response, and measurement errors and can have large impacts on inference.

As an example of mode differences, surveys of risky behaviors such as adolescent sexual behavior, drug use and violence have been studied in Turner et al. (1998). The respondents were randomly assigned to answer surveys by either the traditional questionnaire or a computer-assisted audio interview. It turns out that estimates of those risky behaviors are three times or more higher when computer-assisted audio is used. As another example, a survey of Medicare prostate surgery patients is described in Fowler et al. (1998). Patients in Massachusetts were assigned to mail or personal interviews while patients out of Massachusetts were assigned to mail or telephone. There are 25 statistically significant differences between the telephone and mail responses out of 51 questions compared, but only nine significant differences between mail and personal interviews.

It is often of interest to compare estimates from different surveys of the same population, but methodological differences can make these comparisons useless. “Reconciling” estimates involves estimation procedures that attempt to adjust for the methodological differences and allow comparisons or even combination of data.

J. Van Den Brakel et al. (2020) review different methods to measure discontinuities due to a survey process redesign. For parallel data collection, where data is collected under the old and new designs alongside each other for a certain period, design-based methods in J. A. Van Den Brakel (2008, 2013), state-space models in J. A. Van Den Brakel (2008, 2010) and small area estimation models in Pfeffermann (2002, 2013) and Rao & Molina (2015) can be adopted, depending on the length of the parallel run and sample sizes. For a phase-in approach, where changeover to the new design is done by a gradual roll-out, methods are similar to the parallel run. For the case where there is no overlap at all, state-space models can be used.

Work presented in Chapter 4 of this dissertation develops a small area estimation methodology for reconciling estimates from surveys that have changed over time.

1.2 Existing Methods

Existing methods for testing informative sampling are primarily parametric, but in practice it may be hard to develop parametric models and check the parametric assumptions. In this dissertation, we propose two nonparametric approaches, one based on comparison of empirical distribution functions and one based on maximum mean discrepancy. In Section 1.2.1 and Section 1.2.2, we briefly review nonparametric tests for the classical two-sample problem: given two independent samples, test the hypothesis that they come from the same distribution against the alternative that they do not. These tests motivate our approaches, but in our context, we have a single sample and wish to test the hypothesis that its selection was noninformative against the alternative of informative selection.

Small area estimation models can be used to improve the estimation precision given limited sample size for specific domains. In this dissertation, we adopt small area estimation techniques to adjust for discontinuities between two surveys and reconcile their estimates. Specifically, we propose a Fay–Herriot model with modeled design variances for this problem. As background, we briefly review existing small area estimation methods in Section 1.2.3.

1.2.1 Two-Sample Problems by Empirical Distribution Functions

Random variables X_1, X_2, ... , X_n are independent and identically distributed (iid) if they are mutually independent and they have the same distribution $F_X(x) = \mathbb{P}\left(X_i \leq x\right)$. The random function

$$F_{Xn}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{(X_i \leq x)} \tag{1.2.1}$$

is called the empirical distribution function (edf) of the data, where the indicator function $\mathbb{1}_{(X_i \leq x)} = 1$ when $X_i \leq x$ and 0 otherwise.

The goodness-of-fit problem is one of the classical problems of statistics. It is to test the hypothesis

$$H_0 : F_X(x) = F(x), \tag{1.2.2}$$

where $F(x)$ is a given continuous distribution function.

The Kolmogorov–Smirnov (KS) test and Cramér–von Mises (CvM) test are two of the most widely adopted nonparametric methods for this problem. Kolmogorov (1933) establishes a test based on the statistic

$$K_n = \sqrt{n} \sup_{-\infty < x < \infty} \left| F_{Xn}(x) - F(x) \right|. \tag{1.2.3}$$

If K_n is large, then H_0 in (1.2.2) gets rejected. Denoting the distribution of K_n by $\Phi_n(x)$, Kolmogorov shows that

$$\lim_{n\to\infty} \mathbb{P}\left(K_n \leq x\right) = \lim_{n\to\infty} \Phi_n(x) = \Phi(x) = \sum_{j=-\infty}^{\infty} (-1)^j e^{-2j^2x^2}, \quad 0 < x < \infty. \tag{1.2.4}$$

Cramér (1928) suggests the following criterion:

$$\int_{-\infty}^{\infty} (F_{Xn}(x) - F(x))^2 dK(x),$$

where $K(x)$ is a nondecreasing weight function. Smirnov (1937) modifies this as

$$W_n^2 = n \int_{-\infty}^{\infty} \left(F_{Xn}(x) - F(x)\right)^2 \psi(F(x)) dF(x), \tag{1.2.5}$$

where $\psi(t), 0 \leq t \leq 1$, is a nonnegative weight function. It can be shown that

$$\lim_{n\to\infty} \mathbb{P}\left(W_n^2 \leq x\right) = G(x) = \mathbb{P}\left(\sum_{j=1}^{\infty} \frac{Z_j^2}{\lambda_j} \leq x\right), \tag{1.2.6}$$

where Z_1, Z_2, ... are iid standard normal random variables and λ_1, λ_2, ... are the eigenvalues of the kernel $\Gamma(s,t) = \min\{s,t\} - st$.

The two-sample problem is an extension of the goodness-of-fit problem. Let X_i be as above with continuous distribution function $F_X(x)$. Let $Y_1, Y_2, \ldots, Y_m$ be iid random variables with common continuous distribution function $F_Y(x) = \mathbb{P}\left(Y_i \leq x\right)$ and assume all $n + m$ random variables are mutually independent. Then the two-sample problem is to test the hypothesis

$$H_0 : F_X(x) = F_Y(x). \tag{1.2.7}$$

The edf of Y_i is defined similarly as $F_{Xn}(x)$,

$$F_{Ym}(x) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}_{(Y_i \leq x)}. \tag{1.2.8}$$

Smirnov (1939) establishes the two-sample version of (1.2.3). He shows that the statistic

$$D_{mn} = \sqrt{\frac{mn}{m+n}} \sup_{-\infty<x<\infty} |F_{Xn}(x) - F_{Ym}(x)| \tag{1.2.9}$$

has the same limiting distribution of $\Phi(x)$ as in (1.2.4) if

$$0 < a < \frac{m}{n} < b < \infty, \quad m \to \infty, \quad n \to \infty. \tag{1.2.10}$$

Similarly, by Darling (1957), a natural analogue to (1.2.5) is

$$\frac{mn}{m+n} \int_{-\infty}^{\infty} (F_{Xn}(x) - F_{Ym}(x))^2 \psi \left(\frac{nF_{Xn} + mF_{Ym}}{m+n} \right) d \left(\frac{nF_{Xn} + mF_{Ym}}{m+n} \right). \tag{1.2.11}$$

It has the same limiting distribution as W_n^2 in (1.2.5) when $\psi(t) \equiv 1$.

All of these tests are nonparametric, which means that they do not rely on any parametric distribution assumptions. Our first class of tests are motivated by these nonparametric tests, but replace comparison of edfs from two independent samples by comparison of weighted and unweighted edfs from the same sample. While the mutual independence condition of the classic two-sample problem is not satisfied here, we are still able to establish the asymptotic distribution of our test statistics under reasonable conditions that have been assumed widely in the survey literature.

1.2.2 Maximum Mean Discrepancy

The two-sample problem can be treated as measuring the similarity of two probability measures, and one useful way to measure similarity is to define a probability metric. The characteristics of a distribution are often captured by the integral of some function f with respect to the probability measure p, motivating Müller (1997) to develop the integral probability metrics, $d(p, q)$. For a class of functions $\mathscr{F}$,

$$d(p,q) := \sup_{f \in \mathscr{F}} \left| \int f \, dp - \int f \, dq \right| . \tag{1.2.12}$$

The Kolmogorov–Smirnov test in (1.2.9) is essentially the Kolmogorov metric ρ defined by

$$\rho(p,q) := \sup_{t \in \mathbb{R}} \left| F_X(t) - F_Y(t) \right| , \tag{1.2.13}$$

which is also an integral probability metric, with

$$F_X(t) = \int_X \mathbb{1}_{(-\infty,t]} \, dp,$$

$$F_Y(t) = \int_Y \mathbb{1}_{(-\infty,t]} \, dq,$$

and $\mathscr{F}$ being the set of functions $\mathbb{1}_{(-\infty,t]}, t \in \mathbb{R}$.

The choice of $\mathscr{F}$ is critical. First, $\mathscr{F}$ needs to be large enough so that $d(p, q)$ is a metric. Only in this way can we test $p = q$ by checking $\int f \, dp = \int f \, dq$. Dudley (2002) shows that $p = q$ if and only if $\int f \, dp = \int f \, dq$ for all continuous bounded functions f. However, $\mathscr{F}$ cannot be too large, or else practical use of the metric is infeasible. We work in the setting of finite sample sizes, and require that empirical means converge sufficiently quickly to their expectations. Besides the Kolmogorov metric in (1.2.13), some other examples of $\mathscr{F}$ for which $d(p, q)$ is a metric can be found in Sriperumbudur et al. (2010).

Gretton et al. (2012) and Smola et al. (2007) consider $\mathscr{F}$ to be the unit ball in a universal reproducing kernel Hilbert space (RKHS) $\mathscr{H}$ (Aronszajn (1950)). That is $\mathscr{F}_k = \{f : \|f\|_{\mathscr{H}} \leq 1\}$. The advantages of using $\mathscr{F}_k$ are summarized in Sriperumbudur et al. (2010). An RKHS is defined as follows: $\mathscr{H}$ is a Hilbert space of functions $\mathscr{X} \to \mathbb{R}$ with inner product $\langle \cdot, \cdot \rangle$ satisfying the reproducing property:

$$\langle f(\cdot), k(x, \cdot) \rangle = f(x), \tag{1.2.14}$$

where the reproducing kernel k is assumed to be continuous, symmetric, positive definite and integrable. With the restriction that $\mathscr{X}$ is compact, a universal RKHS is dense in $C(\mathscr{X})$ with respect to the L^∞ norm, where L^∞ norm is defined as $\|f\|_\infty = \sup\{|f(x)| : x \in \mathscr{X}\}$. It follows that

$$\mathbb{E}_{x\sim F_X}\left[f(x)\right] = \mathbb{E}_{x\sim F_X}\left[\langle f(\cdot), k(x,\cdot)\rangle\right] = \langle f(\cdot), \mathbb{E}_{x\sim F_X}\left[k(x,\cdot)\right]\rangle, \tag{1.2.15}$$

and

$$\frac{1}{n}\sum_{i=1}^{n} f(x_i) = \frac{1}{n}\sum_{i=1}^{n}\langle f(\cdot), k(x_i,\cdot)\rangle = \left\langle f(\cdot), \frac{1}{n}\sum_{i=1}^{n} k(x_i,\cdot)\right\rangle. \tag{1.2.16}$$

So we can compute expectations and empirical means by taking inner products with the means in the RKHS. While we have written the above in terms of scalar random variables for simplicity, the random objects can be in more general Hilbert spaces and random vectors are of particular interest in this dissertation.

Gretton et al. (2012) rename the integral probability metric as the Maximum Mean Discrepancy (MMD). Let $\mathscr{F}$ be a class of functions $f : X \to \mathbb{R}$. Then the MMD and its biased empirical estimate are defined as:

$$\mathrm{MMD}[\mathscr{F}, F_X, F_Y] := \sup_{f\in\mathscr{F}}\left(\mathbb{E}_{x\sim F_X}\left[f(x)\right] - \mathbb{E}_{y\sim F_Y}\left[f(y)\right]\right), \tag{1.2.17}$$

and

$$\mathrm{MMD}_b[\mathscr{F}, X, Y] := \sup_{f\in\mathscr{F}}\left(\frac{1}{n}\sum_{i=1}^{n} f(x_i) - \frac{1}{m}\sum_{i=1}^{m} f(y_i)\right), \tag{1.2.18}$$

where $\{x_i\}$ and $\{y_i\}$ are drawn iid from F_X and F_Y respectively.

By restricting $\mathscr{F}$ to be $\mathscr{F}_k$, an unbiased estimator (U-statistic) of MMD is given as

$$\mathrm{MMD}_u^2[\mathscr{F}_k, X, Y] = \frac{1}{n(n-1)}\sum_{i\neq j}^{n}\sum^{n}\widetilde{k}(z_i, z_j), \tag{1.2.19}$$

where $z_i := (x_i, y_i)$ (i.e. assuming $m = n$), and $\widetilde{k}(z_i, z_j) := k(x_i, x_j) + k(y_i, y_j) - k(x_i, y_j) - k(x_j, y_i)$. The asymptotic distribution of $n\mathrm{MMD}_u^2[\mathscr{F}_k, X, Y]$ under the null hypothesis (1.2.7) has been established. Conditions on k such that MMD is a metric are investigated by Fukumizu et al. (2008); Gretton et al. (2007); Smola et al. (2007); Steinwart (2002); Sriperumbudur et al. (2010). Among those, Steinwart (2002) shows that the Gaussian and Laplace kernels are universal. The concentration of empirical means as in (1.2.16) is investigated by Altun & Smola (2006), who show the fast convergence of empirical means to their expectations.

In this dissertation, we adapt the MMD from the two-sample problem to test the hypothesis of informative selection, by comparing weighted and unweighted samples. We establish the asymptotic distribution of our MMD statistic under the null hypothesis of noninformativeness, and investigate the power and size of the test via simulation.

1.2.3 Small Area Estimation

Small area estimation (SAE), also known as small domain estimation, refers to the problem of constructing estimates for small subpopulations using information from a survey sample. As the sample size is limited while the number of domains often substantial, the sample size in each domain is often very small or even zero. However, it is still required to get point estimators with measures of error in these areas. SAE methods can be divided into two categories: "design-based" and "model-based." For design-based methods, the estimator is for some descriptive quantity of the finite population or constructed with the help of a model, but the randomness of the estimator is based on the randomization of the design. Model-based methods, on the other hand, assume a model for the superpopulation and draw inference based on this underlying model. In SAE, because certain areas could have no samples at all, both design-based and model-based methods would have to use auxiliary covariate information. Performance of the model depends heavily on the quality of these auxiliary covariates. A comprehensive review of methods used in SAE is given in Pfeffermann (2013).

Consider a population $U = \{1, \ldots, N\}$, divided into M exclusive and exhaustive areas U_1, U_2, $\ldots$, U_M with N_i units in area i and $\sum_{i=1}^{M} N_i = N$. Let $s = s_1 \cup s_2 \cup \cdots \cup s_m$ be the overall sample, where s_i is the sample of size n_i from sampled area i, and m is the number of sampled areas.

Total sample size is $n = \sum_{i=1}^{m} n_i$. Let y be the variable of interest, and denote y_{ij} the response for unit j in area i, $i = 1, \ldots, M$ and $j = 1, \ldots, N_i$, with sample means $\bar{y}_{ij} = \sum_{j=1}^{n_i} y_{ij}/n_i$. The associated covariate values are $\mathbf{x}_{ij} = (x_{1ij}, \ldots, x_{pij})^{\intercal}$, with sample means $\bar{\mathbf{x}}_i = \sum_{j=1}^{n_i} \mathbf{x}_{ij}/n_i$. The corresponding true area means are $\bar{\mathbf{X}}_i = \sum_{j=1}^{N_i} \mathbf{x}_{ij}/N_i$. We denote the target quantity by θ_i. Suppose the sample is selected by simple random sampling without replacement (SRSWOR) and the targets of interest are area means of the responses, $\theta_i = \bar{Y}_i = \sum_{j=1}^{N_i} y_{ij}/N_i$.

If no covariates are available, the *direct* estimator of θ_i is given by

$$\bar{y}_i = \sum_{j=1}^{n_i} y_{ij}/n_i, \tag{1.2.20}$$

where "direct" means that the estimator uses only the response variable from the sampled area i.

Suppose covariates $\mathbf{x}_{ij}$ are observed with $x_{1ij} \equiv 1$. The *synthetic* estimator is

$$\widehat{\bar{Y}}^{\text{syn}}_{\text{reg},i} = \bar{\mathbf{X}}_i^{\intercal}\widehat{\mathbf{B}} = \frac{1}{N_i}\sum_{j=1}^{N_i} \mathbf{x}_{ij}^{\intercal}\widehat{\mathbf{B}}, \tag{1.2.21}$$

where $\widehat{\mathbf{B}} = (\sum_{i=1}^{m}\sum_{j=1}^{n_i} \mathbf{x}_{ij}\mathbf{x}_{ij}^{\intercal})^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n_i} \mathbf{x}_{ij}y_{ij}$ is the ordinary least square estimators. Here, a common ratio is assumed for all areas between response and covariates. This estimator is *indirect* because it borrows information from other areas, reducing variance relative to the direct estimator. However, this may increase the bias of the estimator, especially if the areas are far from homogeneous in the response.

In order to reduce the possibly large bias, the *survey regression* estimator is given as

$$\begin{aligned} \widehat{\bar{Y}}_i^{\text{S}-\text{R}} &= \bar{\mathbf{X}}_i^{\intercal}\widehat{\mathbf{B}}_w + \frac{1}{N_i}\sum_{j=1}^{N_i} w_{ij}\left(y_{ij} - \mathbf{x}_{ij}^{\intercal}\widehat{\mathbf{B}}_w\right) \\ &= \widehat{\bar{Y}}_{i,\text{HT}} + \left(\bar{\mathbf{X}}_i - \widehat{\bar{\mathbf{X}}}_{i,\text{HT}}\right)^{\intercal}\widehat{\mathbf{B}}_w, \end{aligned} \tag{1.2.22}$$

where $\widehat{\mathbf{B}}_w = (\sum_{i=1}^{m} \sum_{j=1}^{n_i} w_{ij}\mathbf{x}_{ij}\mathbf{x}_{ij}^\intercal)^{-1} \sum_{i=1}^{m} \sum_{j=1}^{n_i} w_{ij}\mathbf{x}_{ij}y_{ij}$. $\{w_{ij}\}$ are sampling weights. $\widehat{\bar{Y}}_{i,\mathrm{HT}}$ and $\widehat{\bar{X}}_{i,\mathrm{HT}}$ are the Horvitz–Thompson (HT) estimators of $\bar{Y}_i$ and $\bar{X}_i$. The survey regression estimator reduces the possibly large bias from the synthetic estimator, but the variance could be large.

The *composite* estimator is constructed to reduce the mean square error, by trading off the low bias/high variance of the survey regression estimator with the high bias/low variance of the synthetic estimator. The composite estimator is defined as

$$\widehat{\bar{Y}}_i^{\mathrm{COM}} = \delta_i \widehat{\bar{Y}}_i^{\mathrm{S-R}} + (1-\delta_i)\widehat{\bar{Y}}_{\mathrm{reg},i}^{\mathrm{syn}}, \quad 0 \le \delta_i \le 1. \tag{1.2.23}$$

With large sample size, the design-based composite estimators are approximately unbiased and consistent. They are protected against model misspecification. However, as the sample size is often small, these estimator can be highly variable. Moreover, they cannot be used to estimate areas with no samples at all.

Model-based methods provide optimal estimators, assuming the correct specification of a model. These methods can overcome disadvantages in design-based models. However, robustness to model misspecification must be investigated.

SAE models are divided into area-level models, in which the model specification is for the direct estimates at the area level, and unit-level models, in which model specification is for the original response data. Usually, the choice of area-level or unit-level is driven by the availability of useful covariates.

One widely used area-level model is by Fay & Herriot (1979). It is defined as:

$$\tilde{y}_i = \mathbf{x}_i^\intercal \boldsymbol{\beta} + u_i + e_i, \tag{1.2.24}$$

where $\tilde{y}_i$ denotes the direct sample estimator of θ_i, u_i are random effects with $u_i \sim \mathcal{N}(0, \sigma_u^2)$ and e_i are the sampling error with $e_i \sim \mathcal{N}(0, \sigma_{Di}^2)$. The random effects u_i are independent from e_i. For known σ_u^2, the best linear unbiased predictor (BLUP) of θ_i is

$$\begin{aligned}\hat{\theta}_i &= \mathbf{x}_i^\intercal \widehat{\boldsymbol{\beta}}_{GLS} + \hat{u}_i \\ &= \mathbf{x}_i^\intercal \widehat{\boldsymbol{\beta}}_{GLS} + \gamma_i \left(\tilde{y}_i - \mathbf{x}_i^\intercal \widehat{\boldsymbol{\beta}}_{GLS} \right) \\ &= \gamma_i \tilde{y}_i + (1-\gamma_i) \mathbf{x}_i^\intercal \widehat{\boldsymbol{\beta}}_{GLS}, \end{aligned} \tag{1.2.25}$$

where $\gamma_i = \sigma_u^2 / \left(\sigma_u^2 + \sigma_{Di}^2\right)$, and $\widehat{\boldsymbol{\beta}}_{GLS} = \left(\mathbf{X}^\intercal \boldsymbol{\Sigma}^{-1} \mathbf{X}\right)^{-1} \mathbf{X}^\intercal \boldsymbol{\Sigma}^{-1} \mathbf{Y}$ is the generalized least square estimator, with $\boldsymbol{\Sigma}$ as the covariance of $\mathbf{Y}$. Fay & Herriot (1979) assume that the design variances σ_{Di}^2 are known, which is not true in practice. Various techniques are used to smooth or otherwise stabilize the estimated design variances before incorporating them into a Fay-Herriot model.

A unit-level model is given by Battese et al. (1988). It is defined as

$$y_{ij} = \mathbf{x}_{ij}^\intercal \boldsymbol{\beta} + u_i + \varepsilon_{ij}, \tag{1.2.26}$$

where $u_i \sim \mathcal{N}(0, \sigma_u^2)$ independent from $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma_\varepsilon^2)$. For known variances σ_u^2 and σ_ε^2, the BLUP of θ_i is

$$\hat{\theta}_i = \gamma_i \left[\bar{y}_i + \left(\bar{\mathbf{X}}_i - \bar{\mathbf{x}}_i \right)^\intercal \widehat{\boldsymbol{\beta}}_{GLS} \right] + (1-\gamma_i) \bar{\mathbf{X}}_i^\intercal \widehat{\boldsymbol{\beta}}_{GLS}, \tag{1.2.27}$$

where $\bar{\mathbf{x}}_i = \sum_{j=1}^{n_i} \mathbf{x}_{ij}/n_i$ and $\gamma_i = \sigma_u^2 / \left(\sigma_u^2 + \sigma_\varepsilon^2/n_i\right)$.

SAE modeling has been extended in many directions, including spatial and temporal extensions, generalized mixed models for discrete responses, multivariate responses, etc. See Cressie (1993), Kim et al. (2001), Opsomer et al. (2003) and Jiang & Lahiri (2006) for review. Another important topic is estimation of mean square error of the SAE predictions, which is challenging even for the linear model. This topic has been investigated, for example, in Kackar & Harville (1984), Harville (1985), Datta & Lahiri (2000), Datta et al. (2005) and Jiang & Lahiri (2006).

In this dissertation, we adopt the Fay-Herriot model as in (1.2.24) to reconcile two sources of surveys and unify them into one mixed model in Chapter 4. We propose a novel model for the design variances and develop a mean squared error approximation.

1.3 Organization

This dissertation addresses two topics in the analysis of complex survey data: nonparametric hypothesis testing for informative selection and using small area estimation to address temporal discontinuities due to survey redesign.

We propose two classes of nonparametric tests for informative selection, each motivated by a nonparametric test for two independent samples. In Chapter 2, we generalize classic two-sample tests that compare empirical cumulative distribution functions, including Kolmogorov–Smirnov and Cramér–von Mises, by comparing weighted and unweighted empirical cumulative distribution functions. In Chapter 3, we adapt two-sample tests that compare distributions based on the maximum mean discrepancy to the setting of weighted and unweighted distributions. The asymptotic distributions of both test statistics are established under the null hypothesis of noninformative selection. Simulation results, for various degrees of informativeness and for different sources of informativeness, demonstrate the usefulness of the asymptotic approximations, and show that our tests have competitive power with parametric tests in a correctly specified parametric setting while achieving greater power in misspecified scenarios. Both nonparametric tests are investigated in applications to data from a recreational fisheries survey.

In Chapter 4, reconciliation of two sets of estimates from two surveys, which differ due to various sources of nonsampling error, are investigated. A log-normal model for the estimates from the two surveys is built. A design variance model is developed to smooth the estimated design variances. The empirical best linear unbiased predictors (EBLUPs) of the reconciled estimates are obtained by applying the modeled design variances in a Fay-Herriot small area estimation model. Mean Square Error (MSE) estimators of EBLUPs are constructed, and empirical evaluation of the MSE estimators is provided in simulations. Results are applied to two surveys in the Marine Recreational Information Program (MRIP), one traditionally collected via telephone interviews, and the other now collected via mailed questionnaires.

A brief discussion of conclusions and future directions is given in Chapter 5. All proofs are deferred to the Appendix.

Chapter 2

Nonparametric Tests for Informative Selection in Complex Surveys

2.1 Introduction

We consider possibly informative selection of a sample from a finite population, with responses Y that are generated as independent and identically distributed (iid) random variables from a cumulative distribution function (cdf) F, referred to as the superpopulation model. The selection is informative if the distribution of the sample responses, given that they were selected, is not iid F. We propose a class of nonparametric procedures to test the null hypothesis of noninformative selection against the alternative of informative selection.

In surveys, the full likelihood of all observable quantities would include not only responses but also design information, such as probabilities of selection (or their inverses, the sampling weights) and response indicators. If the design can be determined to be noninformative, the likelihood of the sample without the design information would be proportional to the full likelihood with the design information, and by the likelihood principle in section 6.3 of Casella & Berger (2002), the inference would be identical. So in the noninformative case, we can ignore the randomness in the design and directly generalize from the sample to the population using likelihood-based procedures. That is, we can apply standard analysis to the sample and get valid inference for the population. If we do incorporate design features in this noninformative setting, we often lose efficiency in the analysis; see, for example, section 3.5 of Chambers et al. (2012). On the other hand, if the design does have informativeness, the analysis must be adjusted to get appropriate results. Failure to account for the informative selection could lead to biased and inconsistent parameter estimators, invalid confidence intervals and errors in conclusions.

Several methods have been established to account for informativeness in analysis. The standard method is Horvitz-Thompson (HT) estimation, in which sampled values are weighted by the inverses of their inclusion probabilities (Horvitz & Thompson (1952)). HT yields unbiased estimators of population totals provided all units have positive probabilities of inclusion and yields consistent estimators of finite population parameters under mild conditions on the sequence of sampling designs. For nonlinear parameters which can be written explicitly as a function of random variables, such as ratios or proportions, the estimator by HT plug-in principle is consistent and asymptotically design-unbiased. If a finite population parameter is the solution to a population-level estimating equation, then the HT plug-in estimator is obtained by solving a weighted sample-level estimating equation. When the parameter is the maximum likelihood estimator (MLE) of the population-level model, then the estimator is the maximum pseudo-likelihood estimator which maximizes the weighted sample-level score function (Binder (1983)). Together, these weighted estimation procedures constitute the standard *design-based* approach to inference. These methods are widely available in software specifically designed for analysis of complex survey data, such as `svymean` in `R` package `survey` and `SURVEYMEANS` procedure in `SAS`.

In contrast to design-based methods, *model-based* methods attempt to describe the joint distribution of the observations, response indicators, and probabilities (full likelihood, Skinner (1994)) or the joint distribution of the observations as distorted by the selection (sample likelihood, Patil & Rao (1978); Breslow & Cain (1988)). Pfeffermann & Sverchkov (1999, 2003) use sample likelihood for the fitting of linear and generalized linear population models. Sverchkov & Pfeffermann (2004) use sample likelihood for the prediction of population totals. Pfeffermann et al. (2006) propose a model-based approach for multi-level modeling under informative multi-stage sampling.

However, methods to adjust inference under informative selection are out of the scope of this chapter, which focuses on testing to determine if the sampling design is informative.

A number of authors have investigated tests of informative selection. One class of tests focuses on whether the moments of the population model residuals are equal to the moments of the sample model residuals. Pfeffermann & Sverchkov (1999) show that the hypothesis of zero correlation between the sampling weights and all order of polynomials of the sample model errors is equivalent to the hypothesis of equal conditional moments for the population model errors and the sample model errors. Pfeffermann & Sverchkov (1999) use standardized Fisher transformation of the correlation to test this hypothesis. This testing procedure is not exact as infinite number of moments should be compared to prove the equivalence of two distributions. In practice, the authors recommend testing the first two to three correlations. Pfeffermann & Sverchkov (1999) show an alternative test that regresses sampling weights against all order of polynomials of sample model errors and test whether the corresponding slopes are zero.

Another important class of tests is based on assessing the significance of the difference between weighted and unweighted estimators of model parameters. DuMouchel & Duncan (1983) construct a test to compare weighted and unweighted parameter estimators in a linear model, and illustrate its equivalence to an $\mathcal{F}$-test of the hypothesis that the original linear model is adequate against the alternative that an augmented linear model (with design matrix expanded to include weighted covariates) is necessary. This equivalence makes the test very convenient to run in standard software. Fuller (1984) considers the case of cluster samples within strata, and gives an approximate $\mathcal{F}$-test. Nordberg (1989) extends the DuMouchel–Duncan test to generalized linear models. Pfeffermann (1993) extends weighted-unweighted comparisons to general likelihood-based problems with explicit estimators, and provides a Wald-type test statistic. Pfeffermann & Sverchkov (2003) extend the test to estimators that are defined as the solutions to estimating equations in generalized linear models.

Our approach builds on the idea of comparing weighted and unweighted estimators. However, instead of assuming parametric models, we propose a class of nonparametric tests of informative selection.

Our results are built on the theory of Boistard et al. (2017a), who establish a functional central limit theorem for the Horvitz-Thompson (HT) empirical process and the Hájek empirical process centered by their finite population mean as well as their superpopulation mean. The results apply to single-stage unequal probability designs and only require conditions on higher-order inclusion probabilities. We apply this theory to establish the asymptotic distribution of the normalized difference between the weighted (Hájek) and unweighted estimators of the cumulative distribution function (cdf). The asymptotic distribution is known up to a scaling constant that can be estimated from the data, so critical values for various test statistics can be obtained. We specifically consider Kolmogorov–Smirnov and Crámer–Von Mises test statistics for informative selection, though other tests could be considered.

This chapter is organized as follows. Notation and assumptions, adopted from Boistard et al. (2017a), are introduced in Section 2.2. Inclusion probabilities up to the fourth order are required and a central limit theorem (CLT) for the Horvitz-Thompson estimator of a population total for independent and identically distributed (iid) bounded random variables is assumed. We prove the asymptotic result for our weighted and unweighted difference in Section 2.3. In Section 2.4, we explore the power and size properties of our tests and compare to existing parametric tests under different types and amounts of informativeness. Lastly, our theory is applied to data from a fisheries survey in Section 2.5. The proofs are deferred to the Appendix.

2.2 Notation and Assumptions

We follow the notation in Boistard et al. (2017a), who adopt the superpopulation setup in Rubin-Bleuer & Kratina (2005).

We first define the superpopulation space. Consider a finite population $U = \{1, \ldots, N\}$. For each element $i \in U$, let $(y_i, z_i) \in \mathbb{R} \times \mathbb{R}_+^q$, with $\mathbf{y} = (y_1, \ldots, y_N) \in \mathbb{R}^N$ denoting the vector of the variable of interest and $\mathbf{z} = (z_1, \ldots, z_N) \in \mathbb{R}_+^{q\times N}$, denoting information about the sampling design. We assume (y_i, z_i) are realizations of random quantities $(Y_i, Z_i) \in \mathbb{R} \times \mathbb{R}_+^q$ that are defined on a common probability space $(\Omega, \mathfrak{F}, \mathbb{P}_m)$.

We next define the design space. Let $\mathcal{S} = \{s : s \subset U\}$ be the collection of subsets of U, and let $\mathfrak{A} = \sigma(\mathcal{S})$ be the σ-algebra generated by $\mathcal{S}$. A sampling design is a function $P : \mathfrak{A} \times \mathbb{R}_+^{q\times N} \mapsto [0, 1]$. Define the probability measure on the design space $(\mathcal{S}, \mathfrak{A})$ as $A \mapsto \mathbb{P}_d(A, \omega) = \sum_{s\in A} P(s, \mathbf{Z}(\omega))$.

Next, we define a product space that includes both the superpopulation space and the design space, assuming conditional independence of sample selection and model characteristics given the design variables. Let $(\mathbf{S} \times \Omega, \mathfrak{A} \times \mathfrak{F})$ be the product space with probability measure $\mathbb{P}_{d,m}$ defined on rectangles $\{s\} \times E \in \mathfrak{A} \times \mathfrak{F}$ by

$$\mathbb{P}_{d,m}(\{s\}, E) = \int_E P(s, \mathbf{Z}(\omega)) \mathrm{d}\mathbb{P}_{\mathrm{m}}(\omega) = \int_E \mathbb{P}_d(\{s\}, \omega) \mathrm{d}\mathbb{P}_{\mathrm{m}}(\omega).$$

For certain sampling designs, the sample size could be random. Denote the sample membership indicator by $\xi_i = 1$ if $i \in s$ and 0 if $i \notin s$, and let $n_s = \sum_{i=1}^{N} \xi_i$ denote the sample size. The first-order inclusion probability is denoted $\pi_i(\omega) = \mathbb{E}_d(\xi_i, \omega) = \sum_{i \in s} P(s, \mathbf{Z}(\omega))$. Further, denote the expected sample size $n = \mathbb{E}_d[n_s(\omega)] = \sum_{i=1}^{N} \mathbb{E}_d(\xi_i, \omega) = \sum_{i=1}^{N} \pi_i(\omega)$. Here $\pi_i(\omega)$ could indicate both the random variable and its realization. For example, the inclusion probability could be a function of $\mathbf{Z}$, i.e. $\pi_i = \pi(Z_i)$.

A functional central limit theorem is obtained by proving weak convergence of all finite dimensional distributions and tightness. To get tightness, we impose a set of conditions involving the sets

$$D_{\nu,N} = \left\{(i_1, i_2, \ldots, i_\nu) \in \{1, 2, \ldots, N\}^\nu : i_1, i_2, \ldots, i_\nu \text{ all different}\right\}, \qquad (2.2.1)$$

for integers $1 \leq \nu \leq 4$. We assume the following conditions:

(C1) There exist constants K_1, K_2, such that for all $i = 1, 2, \ldots, N$,

$$0 < K_1 \leq \frac{N\pi_i}{n} \leq K_2 < \infty, \ \ \omega\text{–a.s.}$$

There exists a constant $K_3 > 0$, such that for all N:

(C2) $\max_{(i,j) \in D_{2,N}} |\mathbb{E}_d(\xi_i - \pi_i)(\xi_j - \pi_j)| \leq K_3 n/N^2$,

(C3) $\max_{(i,j,k) \in D_{3,N}} |\mathbb{E}_d(\xi_i - \pi_i)(\xi_j - \pi_j)(\xi_k - \pi_k)| \leq K_3 n^2/N^3$,

(C4) $\max_{(i,j,k,l) \in D_{4,N}} |\mathbb{E}_d(\xi_i - \pi_i)(\xi_j - \pi_j)(\xi_k - \pi_k)(\xi_l - \pi_l)| \leq K_3 n^2/N^4$,

ω–a.s.

As $N\pi_i/n \leq N/n$, an upper bound in (C1) is immediate if one requires $n/N \to \lambda > 0$. Sometimes, the lower bound is imposed as $\pi_i \geq \pi^* > 0$. This condition can be found in many articles, e.g. Breidt & Opsomer (2000), Bertail et al. (2013), Conti (2014), Conti et al. (2017). Conditions (C2)–(C4) are used to establish tightness of the random processes. These conditions on higher order inclusion probabilities are commonly used in survey sampling. Breidt & Opsomer (2000) show that they hold for simple random sampling without replacement and stratified simple random sampling without replacement, and Boistard et al. (2012) prove that they hold for rejective sampling.

In order to establish weak convergence of all finite dimensional distributions, we need a CLT for suitably-normalized HT estimators of the population means for sequences of bounded iid random variables $V_1, V_2, \ldots$ on $(\Omega, \mathfrak{F}, \mathbb{P}_m)$. Let S_N^2 be the design-based variance of the HT estimator of the population mean,

$$S_N^2 = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}V_iV_j. \tag{2.2.2}$$

We assume the following conditions:

(H1) Let $V_1, V_2, \ldots$ be a sequence of bounded iid random variables, not identical to zero. Suppose there exists an $M > 0$ such that $|V_i| \leq M$ ω–a.s. for all $i = 1, 2, \ldots$, and suppose that for N sufficiently large, $S_N > 0$ and

$$\frac{1}{S_N}\left(\frac{1}{N}\sum_{i=1}^{N}\frac{\xi_i V_i}{\pi_i} - \frac{1}{N}\sum_{i=1}^{N}V_i\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0,1), \qquad \omega\text{–a.s.},$$

where the convergence in distribution is under $\mathbb{P}_d$.

(H2) For $k \in \{1, 2, \ldots\}$, $i = 1, 2, \ldots, N$ and $t_1, t_2, \ldots, t_k \in \mathbb{R}$, define $\mathbf{Y}_{ik}^{*\intercal} = \left(1 - \frac{N\pi_i}{n}\right)\left(\mathbb{1}_{(Y_i \leq t_1)}, \ldots, \mathbb{1}_{(Y_i \leq t_k)}\right)$. There exists a deterministic matrix $\mathbf{\Sigma}_k^*$, such that

$$\lim_{N\to\infty}\frac{n}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}\mathbf{Y}_{ik}^{*}\mathbf{Y}_{jk}^{*\intercal} = \mathbf{\Sigma}_k^*, \quad \omega - \text{a.s.} \tag{2.2.3}$$

(H3) For $k \in \{1, 2, \ldots\}, i = 1, 2, \ldots, N$ and $t_1, t_2, \ldots, t_k \in \mathbb{R}$, define $\mathbf{Y}_{ik}^{\intercal} = \left(1 - \frac{N\pi_i}{n}\right)\left(\mathbb{1}_{(Y_i \le t_1)} - F(t_1), \ldots, \mathbb{1}_{(Y_i \le t_k)} - F(t_k)\right)$. There exists a deterministic matrix $\mathbf{\Sigma}_k$, such that

$$\lim_{N \to \infty} \frac{n}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \mathbf{Y}_{ik} \mathbf{Y}_{jk}^{\intercal} = \mathbf{\Sigma}_k, \quad \omega - \text{a.s.} \tag{2.2.4}$$

(H4) $n/N \to \lambda \in (0, 1)$.

(H5) For $k \in \{1, 2, \ldots\}$ and $t_1, t_2, \ldots, t_k \in \mathbb{R}$, the matrix $\mathbf{\Sigma}_k$ in (2.2.4) is positive definite.

Condition (H1) has been checked by many authors in different sampling designs. It holds for simple random sampling without replacement if $n(N-n)/N \to \infty$ as $N \to \infty$ (see Thompson (2013)), and Poisson sampling under certain conditions (see Fuller (2009)). Hájek (1964) gives a condition that is sufficient and necessary for (H1) for rejective sampling. Berger (1998) extends it to high entropy designs. Conditions like (H2) and (H3) are quite standard in survey sampling literature. See, for example Deville & Särndal (1992), or Francisco & Fuller (1991). Condition (H4) is also quite standard; see, for example Breidt & Opsomer (2000) or Conti (2014).

2.3 Asymptotic Results

2.3.1 Asymptotic Distribution

Theorem 2.3.1. *Let $Y_1, \ldots, Y_N$ be iid random variables with cdf F. Suppose that conditions (C1)–(C4), (H1)–(H5) hold, then*

$$\mathbb{T}_n = \sqrt{n}\left(\widehat{F}_{HJ} - \widehat{F}\right) = \sqrt{n}\left(\frac{1}{\widehat{N}} \sum_{i=1}^{N} \frac{\xi_i}{\pi_i} \mathbb{1}_{(Y_i \le t)} - \frac{1}{n} \sum_{i=1}^{N} \xi_i \mathbb{1}_{(Y_i \le t)}\right)$$

converges weakly to a mean zero Gaussian process $\mathbb{G}$ with covariance function

$$\mathbb{E}_m \mathbb{G}(s)\mathbb{G}(t) =$$

$$\lim_{N\to\infty} \frac{n}{N^2} \sum_{i=1}^{N}\sum_{j=1}^{N} \mathbb{E}_m \left[\frac{\pi_{ij}}{\pi_i \pi_j} \left(1 - \frac{N\pi_i}{n}\right)\left(1 - \frac{N\pi_j}{n}\right) \left(\mathbb{1}_{(Y_i \le s)} - F(s)\right)\left(\mathbb{1}_{(Y_j \le t)} - F(t)\right)\right]$$

for $s, t \in \mathbb{R}$, where $\widehat{N} = \sum_{i=1}^{N} \xi_i/\pi_i$.

Corollary 2.3.1.1. *Under the conditions of Theorem 2.3.1, if the Y_i's are independent of π_{ij}'s, for all i and j, then the covariance function can be simplified as*

$$\mathbb{E}_m \mathbb{G}(s)\mathbb{G}(t) = \lim_{N\to\infty} \frac{n}{N^2} \sum_{i=1}^{N} \mathbb{E}_m \left[\frac{1}{\pi_i} \left(1 - \frac{N\pi_i}{n}\right)^2 \right] [F(\min\{s,t\}) - F(s)F(t)] \textit{ for } s, t \in \mathbb{R}.$$

Corollary 2.3.1.2. *Consider probability proportional to size (pps) sampling with $\pi_i = nz_i / \sum_U z_i$, where the z_i's are iid with $\mu_z = \mathbb{E}_m[z_1] < \infty$ and $\mathbb{E}_m [1/z_1] < \infty$. Then, under the conditions of Corollary 2.3.1.1, the covariance function can be further simplified as*

$$\mathbb{E}_m \mathbb{G}(s)\mathbb{G}(t) = \left(\mu_z \mathbb{E}_m \left[\frac{1}{z_1}\right] - 1\right) [F(\min\{s,t\}) - F(s)F(t)] \textit{ for } s, t \in \mathbb{R}.$$

2.3.2 Test Statistic

The classical functional central limit theorem tells us that the Brownian Bridge of $F(t)$, denoted $\mathbb{B}\left(F(t)\right)$, is a Gaussian process with zero mean and covariance function $[F(\min\{s,t\}) - F(s)F(t)]$. Further, $\|\mathbb{B}\left(F(t)\right)\|_\infty$ follows the Kolmogorov distribution (Donsker (1951)) and $\int_t \mathbb{B}\left(F(t)\right)^2 dF(t)$ follows the Cramér–von Mises distribution (Donsker (1951)), where $\|\cdot\|_\infty$ is the sup norm. Then by Corollary 2.3.1.1, we can construct the test statistics in the following result.

Result 2.3.1.1. *Under conditions in Theorem 2.3.1 and the null hypothesis of noninformative selection, $\mathbb{T}_n(t) \xrightarrow{\mathcal{L}} C^{1/2}\mathbb{B}\left(F(t)\right)$, where the scaling constant*

$$C = \lim_{N\to\infty} \frac{n}{N^2} \sum_{i=1}^{N} \mathbb{E}_m \left[\frac{1}{\pi_i} \left(1 - \frac{N\pi_i}{n}\right)^2 \right]$$

can be consistently estimated by

$$\widehat{C} = \frac{n}{\widehat{N}^2} \sum_{i \in U} \frac{\xi_i}{\pi_i^2} \left(1 - \frac{\widehat{N}\pi_i}{n}\right)^2 = \frac{n-1}{n} \left(\frac{S_w}{\bar{w}}\right)^2, \tag{2.3.1}$$

with $\bar{w}$ the sample mean and S_w the sample standard deviation of the design weights (inverse inclusion probabilities). Then, asymptotically, $\widehat{C}^{-\frac{1}{2}} \left\|\mathbb{T}_n(t)\right\|_\infty$ follows the Kolmogorov distribution and $\widehat{C}^{-1} \int \mathbb{T}_n^2(t) d\widehat{F}(t)$ follows the Cramér–von Mises distribution.

Remark 2.3.1.1. *Consider the special case of Poisson sampling, under the null hypothesis of noninformative selection with iid $\{\pi_i\}$,*

$$\begin{aligned}\mathbb{T}_n &= \frac{\sqrt{n}}{\widehat{N}} \sum_{i=1}^{N} \frac{\xi_i}{\pi_i} \left(1 - \frac{\widehat{N}\pi_i}{n}\right) \left(\mathbb{1}_{(Y_i \leq t)} - F(t)\right) \\ &= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \sqrt{\frac{n}{N}} \frac{\xi_i}{\pi_i} \left(1 - \frac{N\pi_i}{n}\right) \left(\mathbb{1}_{(Y_i \leq t)} - F(t)\right) + o_p(1),\end{aligned}$$

so the asymptotic distribution of $\mathbb{T}_n$ in Result 2.3.1.1 above can also be derived by directly applying Theorem 2.9.2 (the multiplier central limit theorem) of van der Vaart & Wellner (1996). Theorem 2.3.1 covers more general designs and asymptotic distributions under alternative hypotheses.

2.4 Simulation

2.4.1 Gestational Age Example

As described in Bonnéry et al. (2018), this example simulates sampling of at-risk infants, motivated by the actual design of the 1988 National Maternal and Infant Health Survey (NMIHS). That study, described in Korn & Graubard (1999), oversampled low birthweight infants. Unlike the original setting of a stratified sampling design, we simulate an unstratified design with $N = 15000$ and $n = 300$. The variable of interest is $Y =$ gestational age, assumed to be iid $\mathcal{N}(\theta, \sigma^2)$ in the superpopulation. As the survey is designed to oversample the lower birthweight infants, the inclusion probabilities (and hence the sampling weights) depend directly on birthweight, which in turn is highly correlated with gestational age. Hence, the design is informative.

We generate the inclusion probabilities by

$$\ln \pi \mid (Y = y) \sim \mathcal{N}(-\delta_0 - \delta y, \tau^2)$$

with

$$\delta_0 = -\ln \frac{n}{N} + \frac{\tau^2}{2} - \delta\theta + \frac{\delta^2\sigma^2}{2}.$$

We set $\theta = 39.853$, $\sigma^2 = 16.723$, $\tau^2 = 0.087$ and let δ range from 0 to 0.03 with grid size of 0.002. Lower values of δ correspond to less informativeness, with noninformativeness when $\delta = 0$.

For each of the 1000 independent replicates, we draw a Poisson sample of expected size $n = 300$. Then we compute $\mathbb{T}_n$'s as in Theorem 2.3.1, estimate the scaling factor C's as in (2.3.1) and calculate the Kolmogorov–Smirnov (KS) and Cramér-von Mises (CvM) statistics.

Figure 2.1 shows the power of our methods compared to DuMouchel and Duncan (DD, DuMouchel & Duncan (1983)) and Pfeffermann (PFE, Pfeffermann (1993)) versus δ; informativeness increases with δ. All methods maintain approximately the correct size at the noninformative null, $\delta = 0$. In this example, our nonparametric methods show competitive power relative to DD and PFE. DD has the correct parametric model specification and PFE has the correct likelihood. The KS statistic shows lowest power while CvM power curve is slightly below DD. PFE power curve lies between CvM and KS. The reason PFE does not perform as well as DD is that we estimate the variance in the likelihood instead of treating it as known. In other simulations (not shown), if we plug in the true variance in the likelihood, the PFE power curve lies almost right on top of DD power curve. Here DD assumes correct model and PFE uses correct likelihood, while our methods do not assume parametric distributions but only use sample responses and their design weights. Essentially, our methods achieve good power for "free"; that is, without any of the modeling cost of the parametric methods.

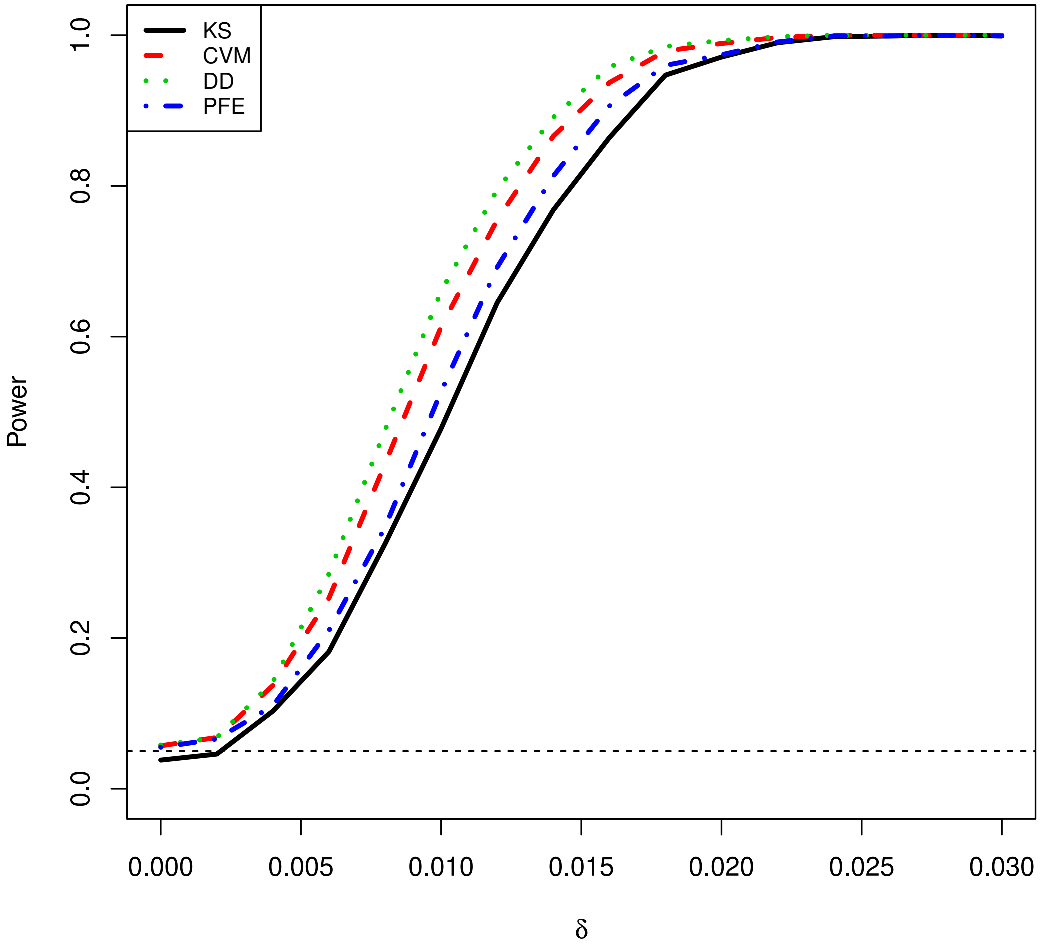

Figure 2.1: Power versus informativeness for DuMouchel–Duncan (DD; top curve), Pfeffermann (PFE; second lowest curve), Kolmogorov–Smirnov (KS; bottom curve), and Cramér–von Mises (CvM; second highest curve) tests, based on 1000 replicate Poisson samples of expected size $n = 300$ from the simulated gestational age population. Nominal size of all tests is $\alpha = 0.05$ (horizontal reference line). Informativeness increases with the value of δ, with $\delta = 0$ the noninformative null.

2.4.2 Two-Stage Gestational Age Example

The theory of Section 2.3.1 is derived for single-stage sampling. We investigate the application to two-stage sampling via simulation, by extending the gestational age example of Section 2.4.1. Consider a finite population consisting of $N_{PSU} = 15000$ primary sampling units (PSUs), each made up of $m = 10$ secondary sampling units (SSUs). We set Z_{kl} iid $\mathcal{N}(\theta/m, \sigma^2/m)$ for $k = 1, 2, \ldots, N_{PSU}$ and $l = 1, 2, \ldots, m$, so that the PSU totals $Y_k = \sum_{l=1}^{m} Z_{kl}$ are iid $\mathcal{N}(\theta, \sigma^2)$ across PSUs. We generate the inclusion probabilities of the first stage by

$$\ln \pi \mid Y = y \sim \mathcal{N}(-\delta_0 - \delta y, \tau^2)$$

with

$$\delta_0 = -\ln \frac{n}{N} + \frac{\tau^2}{2} - \delta\theta + \frac{\delta^2\sigma^2}{2}.$$

We set $\theta = 39.853$, $\sigma^2 = 16.723$, $\tau^2 = 0.087$ and δ ranges from 0 to 0.03 with grid size of 0.002 as before. For each of 1000 independent replicates, we draw a Poisson sample of the PSUs with expected sample size $n = 300$, and then draw a simple random sample without replacement (SRSWOR) within each selected PSU, with sample size n_m.

The goal in this problem is inference for the distribution of Y, the PSU totals. To test for informative selection, we compute the estimated PSU totals, $\widehat{y}_k = m \sum_{l \in s_k} z_{kl}/n_m$, from the simple random sample s_k we collected in the second stage. We then follow the calculations as in Section 2.4.1, using estimated PSU totals $\widehat{y}_k$ in place of y_k and using PSU weights w_k. Figure 2.2 shows the power curves of all the statistics for various amount of informativeness with $n_m = 2$ and $n_m = 6$. The setting in Section 2.4.1 can be treated as a special case of two-stage sampling with $n_m = 10$. In spite of the subsampling, the tests maintain the correct size and the same relative power ordering as in Figure 2.1, though power decreases as the subsample size n_m decreases.

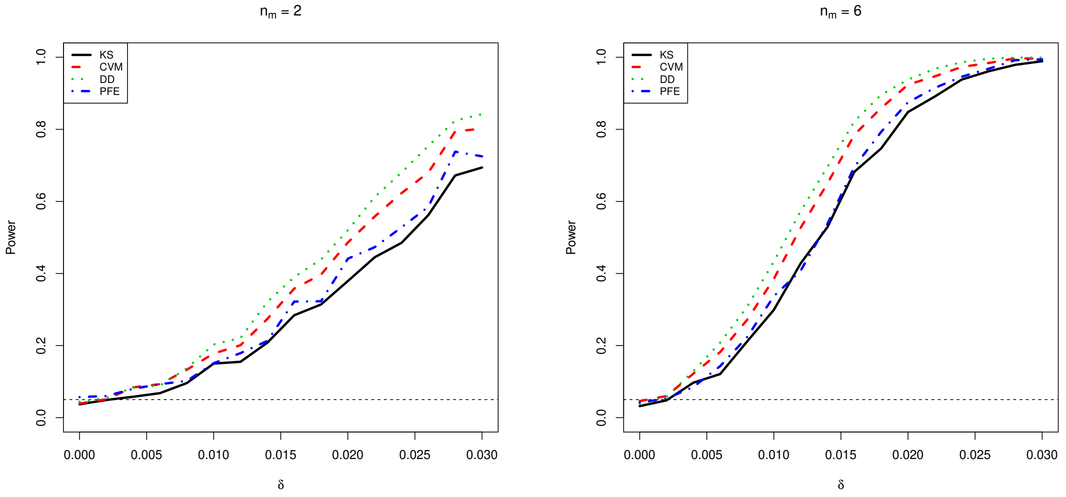

Figure 2.2: Power versus informativeness for DuMouchel–Duncan (DD; top curve), Pfeffermann (PFE; second lowest curve), Kolmogorov–Smirnov (KS; bottom curve), and Cramér–von Mises (CvM; second highest curve) tests, based on 1000 replicate two-stage samples, with Poisson samples of expected size $n = 300$ PSUs in stage one, and simple random samples without replacement of size $n_m = 2$ (left panel) and $n_m = 6$ (right panel) in stage two. Nominal size of all tests is $\alpha = 0.05$ (horizontal reference line). Informativeness increases with the value of δ, with $\delta = 0$ the noninformative null.

2.4.3 Stratified Simple Random Sampling without Replacement

Theorem 2.3.1 is derived in the sense that the y's are assumed iid from the superpopulation. It is of interest to study the case in which y-properties vary across strata. We consider the example of stratified simple random sampling, with strata formed by ordering on an auxiliary variable z that might be correlated with y, so that the distribution of y varies across strata except in the case of no correlation between y and z. We generate the finite population by first simulating $H = 100$ random stratum sizes $N_h \sim$ Negative Binomial for $h = 1, 2, \ldots, H$, with mean 150 and number of success trials 10. Here the probability mass function of the Negative Binomial is $\Pr[X = k] = \binom{k+r-1}{k} p^r (1-p)^k$, where r is the number of successes, k is the number of failures, and p is the probability of success. The population size is then $N = \sum_{h=1}^{H} N_h$, with expected population size $H\mathbb{E}[N_h] = 15000$. We generate N iid random vectors

$$\begin{pmatrix} y \\ z \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \theta \\ \theta \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{pmatrix} \right)$$

and sort (y, z) by z in ascending order: the first N_1 pairs are then stratum 1, the next N_2 are stratum 2, etc. Next, we compute the sum of z in each stratum, t_{zh}, and allocate the stratum sample size n_h proportional to this total: $n_h = \lfloor n * t_{zh} / \sum_{h=1}^{H} t_{zh} \rceil$. We then select the sample by stratified simple random sampling without replacement.

We apply our test at the stratum level instead of the observation level, comparing the unweighted empirical cdf of the H stratum sample means $\bar{y}_h = n_h^{-1} \sum_{k \in s_h} y_k$ to the weighted empirical cdf using stratum weights $w_h = N_h n_h^{-1}$. We conjecture that our asymptotic framework could be adapted to this setting, assuming that the number of strata goes to infinity while the size of each stratum shrinks, so that the empirical cdf of the stratum population means converges to the superpopulation cdf of y. Our simulation results support this conjecture.

We set $\theta = 40$, $\sigma = 0.5$ and let ρ vary from 0 to 1 with grid size 0.025, where $\rho = 0$ is noninformative and informativeness grows as ρ increases.

Figure 2.3 shows the power curves of all the statistics versus ρ. All tests have approximately the correct test size at $\rho = 0$ when the critical value is determined from our asymptotic theory with the number of strata, H, replacing the expected sample size, n. PFE statistic shows highest power, with DD statistic slightly under PFE and CvM slightly under DD. KS statistic shows lowest power. PFE shows better performance than it does in the gestational age example, because the informativeness affects both the mean and the variance in this setting, and the likelihood-based method is able to detect these differences. Once again, the nonparametric methods have considerable power to detect informative selection without any need for a correct parametric specification.

2.4.4 Scaled *t* Distribution

In the gestational age example, both DD and PFE show good power as expected: the informativeness exists in the mean and the correctly-specified $\mathcal{F}$-test and Wald test should be able to capture the difference. In this t example, the informativeness will appear in the variance.

Our variable of interest y_k is generated as follows:

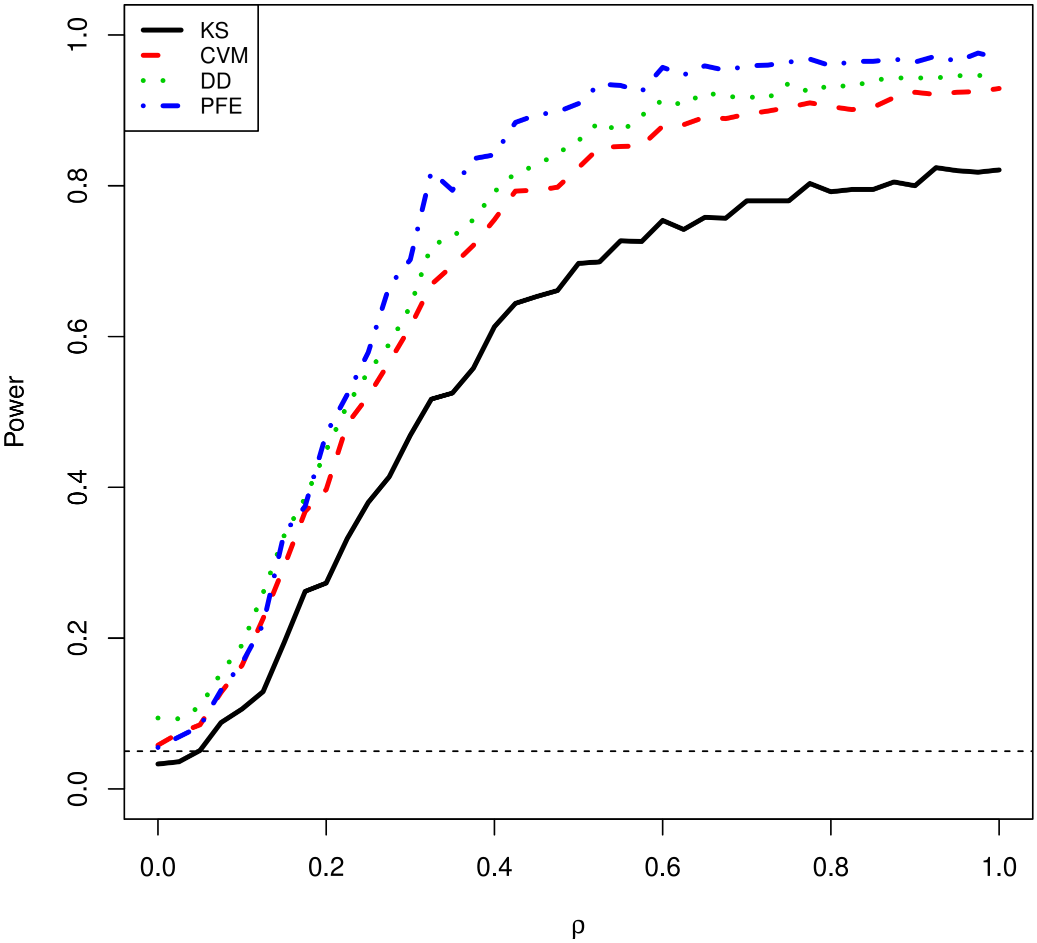

Figure 2.3: Power versus informativeness for Pfeffermann (PFE; top curve), DuMouchel–Duncan (DD; second highest curve), Kolmogorov–Smirnov (KS; bottom curve) and Cramér–von Mises (CvM; second lowest curve) tests, based on 1000 replicate stratified simple random samples without replacement of size $n = 300$ from the simulated bivariate normal distribution. Nominal size of all tests is $\alpha = 0.05$ (horizontal reference line). Informativeness increases with the value of ρ, with $\rho = 0$ the noninformative null.

$$y_k = \mu + \sigma \frac{z_k}{\sqrt{v_k/\nu}} \sqrt{\frac{\nu - 2}{\nu}} = \mu + \sigma_k z_k,$$

where $\{z_k\}$ is iid $\mathcal{N}(0,1)$, independent of $\{v_k\}$ iid χ^2_ν. The error terms here are distributed as scaled t_ν, with mean 0 and variance σ^2 for $\nu > 2$, and

$$y_k \mid v_k \sim \mathcal{N}(0, \sigma_k^2).$$

Let $\{v_k^*\}$ be iid χ^2_ν, independent of $\{z_k\}$ and $\{v_k\}$, and set

$$\tau_k = \sigma \frac{1}{\sqrt{v_k^*/\nu}} \sqrt{\frac{\nu - 2}{\nu}},$$

which has the same distribution as σ_k but is not used in generating y_k. Define $d_k = \rho\sigma_k + (1-\rho)\tau_k$ and set the inclusion probabilities as

$$\pi_k = n d_k \left(\sum_{k \in U} d_k \right)^{-1},$$

where $\rho \in [0,1]$ is a constant to control the amount of informativeness in the design. We select samples by Poisson sampling with probability proportional to size d_k. The motivation for this example is that designs with $\pi_k \propto \sigma_k$ minimize the unconditional variance, with respect to model and design, of the Horvitz-Thompson estimator of the y-total, and d_k is a proxy for σ_k.

We set $\mu = 39.853$, $\sigma^2 = 6.123$, $\nu = 5$, and let ρ range from the noninformative null at $\rho = 0$ to the highly-informative, optimal design at $\rho = 1$ with grid size of 0.025. For each of 1000 independent replicates, we draw a Poisson sample of expected size $n = 300$. As in the gestational age example, we compute $\mathbb{T}_n$'s, estimate C's and calculate the KS and CvM statistics.

Figure 2.4 shows the power of our methods compared to DD and PFE under various amounts of informativeness. All tests have approximately the correct size at the noninformative null, $\rho = 0$. As ρ increases, DD has a very low amount of power, because the weighted and unweighted estimates of the mean will differ by chance, and the DD test will correctly reject the null by incorrectly attributing the difference to bias caused by informative selection: both the weighted and unweighted estimates are actually unbiased for the mean. PFE uses the correctly-specified likelihood and has the most power. Our nonparametric tests are much less powerful than PFE, but again these methods use only the sample observations and weights, requiring no parametric modeling at all, and thus are always worth trying.

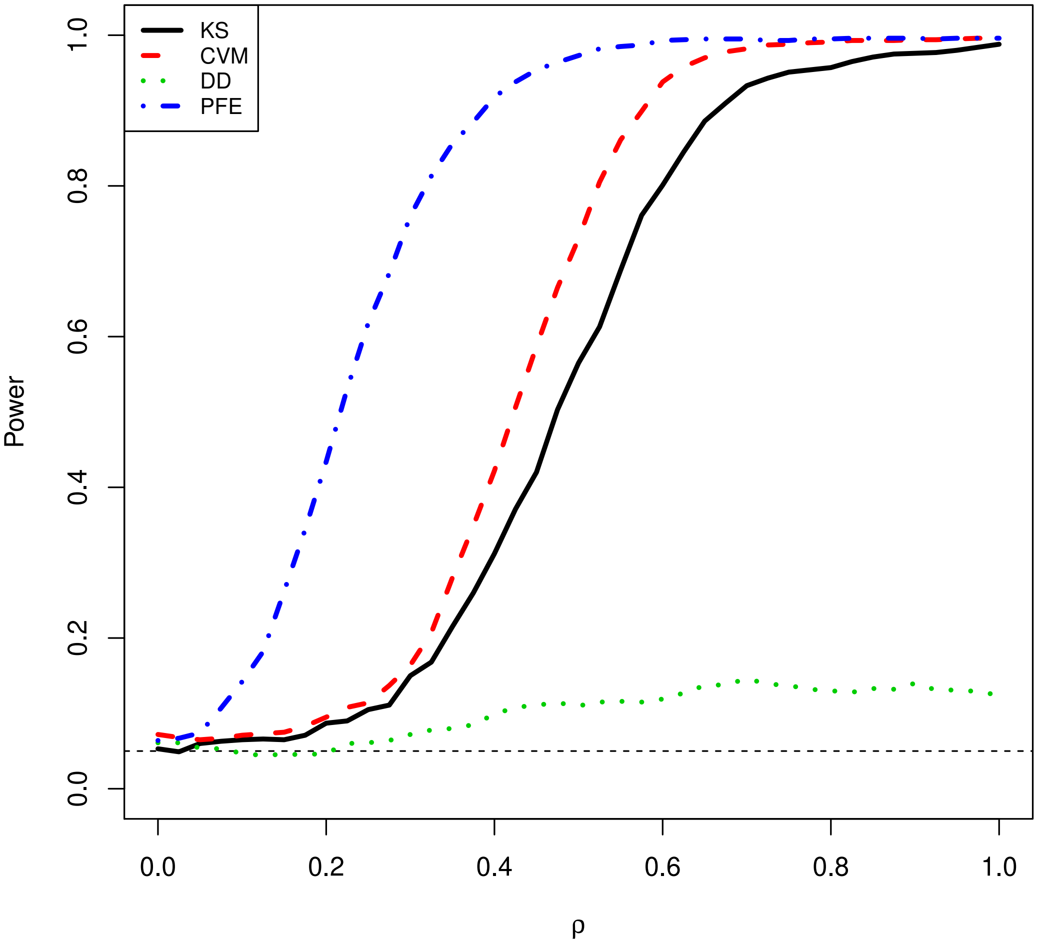

Figure 2.4: Power versus informativeness for DuMouchel–Duncan (DD; bottom curve), Pfeffermann (PFE; top curve), Kolmogorov–Smirnov (KS; second-lowest curve), and Cramér–von Mises (CvM; second highest curve) tests, based on 1000 replicate Poisson samples of expected size $n = 300$ from the simulated scaled-t population. Nominal size of all tests is $\alpha = 0.05$ (horizontal reference line). Informativeness increases with the value of ρ, with $\rho = 0$ the noninformative null.

2.4.5 Pretest Estimators for Scaled t Distribution

As mentioned in Section 2.1, if the sampling design is noninformative, we should use the unweighted estimator for efficiency. Otherwise, we should use a design-weighted estimator or other estimation methods that incorporate design information.

The availability of our test suggests an alternative, "pretest estimator": use the unweighted estimator if the test fails to reject the null hypothesis of noninformativeness, or use the weighted estimator if the test rejects the null.

We compared the root mean squared error (rMSE) of the weighted, unweighted and pretest estimators (using KS or CvM as the test) via simulation for the mean, median, upper quartile and 90th percentile of the scaled t distribution as in Section 2.4.4. Figure 2.5 shows the rMSE of these estimators as a function of the amount of informativeness. In each panel, the dotted line shows the rMSE of the unweighted estimator, which increases due to increasing variability of the sample as ρ increases, and due to bias under informative selection for the quantiles (but not the mean). The dashed line sloping down from the left shows the rMSE of the weighted estimator, which loses some efficiency due to unnecessary weighting when ρ is small, but reduces bias without an increasing price in variance as ρ increases. The other two curves, solid for KS and dash-dot for CvM, show the rMSE of the pretest estimators.

For the mean, shown in the upper left panel of Figure 2.5, the weighted estimator does not lose much efficiency in low informativeness, and both pretest estimators have large variance in moderate informativeness. So the weighted estimator is the best one for estimating the mean. Figure on the upper right is for the median. With the increase in ρ, the unweighted estimator has huge bias and thus large rMSE, while the weighted estimator loses efficiency in lower ρ, but stays consistent overall. Our two pretest estimators have both advantages of efficiency in low informativeness and unbiasedness in high informativeness. The KS and CvM pretest estimators behave similarly, each dominating the weighted estimator for low values of ρ, dominating the unweighted estimator for intermediate values of ρ, and converging to the rMSE of the weighted estimator for high values of ρ. The lower left and lower right are figures for the upper quartile and 90th percentile. Both pretest estimators gain efficiency in low informativeness while having some bias for intermediate ρ, and produce similar rMSE overall as the weighted estimator. To sum up, our pretest estimators could have better performance than the weighted estimator for certain problems.

2.5 Application to a Recreational Angling Survey

We apply the nonparametric tests of informative selection to recreational angling data from the 2016 Marine Recreational Information Program (MRIP) in South Carolina. MRIP measures the number of fishing trips taken by recreational anglers in saltwater, along with the number of fish of each species caught by the anglers. Data on recreational angling are important for understanding fish stocks and sustainable management of fisheries.

Both shore fishing and boat fishing are of interest in MRIP. In this example, we focus on boat trips. To estimate characteristics of the population of all boat trips in South Carolina in 2016, MRIP uses two complementary surveys: an on-site "intercept" survey that collects catch by species information via angler interviews at the fishing site, and an off-site "effort" survey that collects information on the number of angler trips via self-administered questionnaire (mail-out/mail-back). The two sources of information are combined into the weights for the intercept survey, so we focus here on the details of that survey.

In MRIP, the intercept survey of boat trips is obtained by constructing a frame of publicly-accessible sites where boats can return to shore; crossing those sites with days in the fishing season to get "site-days"; stratifying site-days spatially (using contiguous South Carolina counties as strata) and temporally (using five two-month waves: March–April,. . . , November–December); obtaining a stratified sample of site-days; and intercepting all boat trips on selected site-days. The stratified sampling is conducted with probabilities proportional to estimated fishing activity for the site-days ("pressures"). The weights for MRIP reflect these unequal probabilities of selection and also reflect other adjustments, particularly from the effort survey.

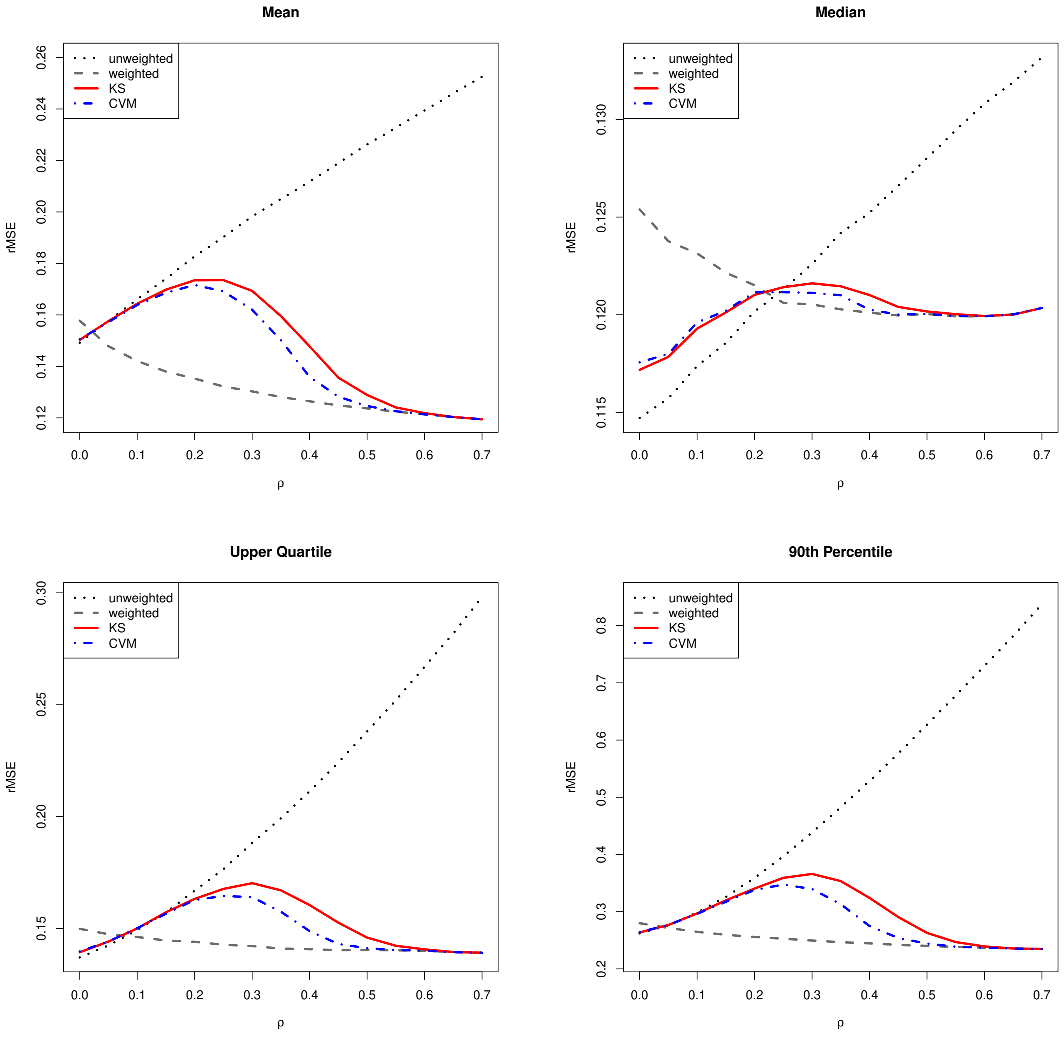

Figure 2.5: rMSE of unweighted estimator (increasing dotted curve), weighted estimator (decreasing dashed curve) and pretest estimators with KS statistic (solid curve) and CvM statistic (dashed-dotted curve) for the mean (upper left panel), median (upper right panel), upper quartile (lower left panel) and 90th percentile (lower right panel). Results based on 1000 replicate Poisson samples of expected size $n = 300$ from the simulated scaled t population. Nominal size of tests for pretest estimators is $\alpha = 0.05$. Informativeness increases with the value of ρ, with $\rho = 0$ the noninformative null.

Table 2.1: p-values of Kolmogorov–Smirnov (KS), Cramér–von Mises (CvM) and Dumouchel–Duncan (DD) statistics for eight different response variables: number of anglers and catch for seven species.

Variable of Interest	KS	CvM	DD
Anglers	0.020	0.007	0.027
Red drum	0.998	0.930	0.763
Black sea bass	0.095	0.091	0.060
Bluefish	0.650	0.569	0.853
Black drum	0.585	0.603	0.836
Wahoo	0.895	0.703	0.575
Gag grouper.	0.815	0.655	0.300
Atlantic croaker	0.719	0.637	0.388

Table 2.1 shows the p-values of KS, CvM and DD statistics for various variables. At test size $\alpha = 0.05$, all of the tests give same result. The only significant variable is Anglers. From the nature of the design, we know Anglers ought to be informative, as surveys were taken on the days and sites that were more likely to have anglers. After investigating the plots of Anglers versus catch for the seven species, we do not see strong relationships, and conclude that it is reasonable to treat the design as noninformative for the other variables.

2.6 Discussion

This chapter provides nonparametric methods of testing for informative selection by following the idea of one class of parametric tests that compare weighted and unweighted estimators of parameters. The test statistics turn out to be the scaled version of well-known KS and CvM statistics. Relatively good power, being free from distribution assumptions and low computational cost make this method always worth trying if possible. However, it should be clear that this method does not work for every survey design. The result is developed under the framework of single-stage unstratified design, including poisson sampling, rejective sampling and high entropy designs. Heuristically, it performs similarly under two-stage sampling and stratified sampling as shown in the simulations. To extend the current results, theories under multistage designs and stratified sampling deserve attention.

Chapter 3

Nonparametric Tests for Informative Selection by Maximum Mean Discrepancy

3.1 Introduction

Let Y be a random variable in Hilbert space and let q denote its probability measure, also known as the superpopulation model. Our examples are Euclidean spaces, with real-valued scalar random variables or real random vectors, but the methodology we describe is not limited to these cases. Responses for a finite population of N elements are generated as independent and identically distributed (iid) realization of Y. We consider possibly informative selection of a sample from this finite population. The selection is informative if the conditional probability measure of the sampled responses, given that they were selected, is no longer iid q. We establish a nonparametric procedure based on the maximum mean discrepancy to test the null hypothesis of noninformative sampling versus the alternative hypothesis of informative sampling.

If the sampling design is noninformative, we can ignore the randomness in the design and directly draw inferences based on the sample using classical likelihood-based procedures. That is, standard inferences based on the sample are still valid for the population. Otherwise we lose efficiency by incorporating the design into the analysis; see section 3.5 of Chambers et al. (2012). On the other hand, if there is evidence that the design is informative, we must take into account the informative selection to get valid inferences. Failing to do so can produce biased and inconsistent parameter estimators, false confidence intervals and erroneous conclusions.

There are two major approaches to adjust for the effect of informative selection. The standard approach is *design-based* estimation, which builds on Horvitz-Thompson (HT) estimation. Sampled values are weighted by the inverses of their inclusion probabilities to get unbiased estimators of population totals, assuming all units have positive inclusion probabilities (Horvitz & Thompson (1952)). HT estimators are consistent estimators of finite population parameters under mild conditions on the sequence of sampling designs. For nonlinear parameters such as ratios or proportions, which are explicit functions of finite population totals, the estimators obtained by plugging in HT estimators are consistent and asymptotically design-unbiased under mild conditions. For a parameter that is a solution to a finite population-level estimating equation (expressed as a finite population total), the HT plug-in estimator is the solution to the weighted sample-level estimating equation, and is again consistent and asymptotically design-unbiased under mild conditions. One important example is estimation of the finite population parameter defined as the maximum likelihood estimator (MLE) at the finite population-level; that is, the parameter is the solution of the finite population-level score equation. The corresponding estimator is the maximum pseudo-likelihood estimator, obtained by maximizing the weighted log-likelihood, or equivalently solving the weighted sample-level score equation (Binder (1983)). These weighted estimation procedures together constitute the standard *design-based* approach to inference. These methods are widely adopted in software specifically designed for survey analysis, such as the `R` package `survey` and `SURVEYMEANS`, `SURVEYLOGISTIC`, etc. procedures in `SAS`.

Another approach is the class of *model-based* methods, which attempt to describe the joint distribution of the observations as distorted by the selection (sample likelihood: Patil & Rao (1978); Breslow & Cain (1988)) or the joint distribution of the observations, response indicators, and probabilities (full likelihood: Skinner (1994)). Pfeffermann & Sverchkov (1999, 2003) fit linear and generalized linear population models by sample likelihood. Sverchkov & Pfeffermann (2004) predict population totals by sample likelihood. Pfeffermann et al. (2006) propose a multi-level model under informative multi-stage sampling.

However, estimation methods to account for informative selection are out of the scope of this chapter. We focus on testing to determine if the sampling design is informative.

Several authors have investigated hypothesis test of informative selection. One class of tests checks whether all moments of residuals of the population model are identical to the moments of residuals of the sample model. Pfeffermann & Sverchkov (1999) show that the hypothesis of equal moments of residuals for the population model and the sample model is equivalent to the hypothesis of zero correlation between the sampling weights and all orders of polynomials of residuals of the sample model. The hypothesis is tested by comparing the standardized Fisher transformation of the correlation to the mean zero normal distribution under the null of noninformative selection. This testing procedure is, however, not exact as (theoretically) an infinite number of correlations between polynomials of residuals and sampling weights should be compared. The authors recommend testing the first two to three correlations in practice. An alternative test is also provided by Pfeffermann & Sverchkov (1999), which regresses sampling weights against all polynomials of sample model residuals and tests whether the corresponding slopes are zero.

Another major class of tests focuses on evaluating whether the difference between weighted and unweighted estimators of model parameters is significant. A test is constructed by DuMouchel & Duncan (1983) to compare weighted and unweighted parameter estimators in a linear model, and it is equivalent to an $\mathcal{F}$-test of the null hypothesis that the original linear model is adequate versus the alternative hypothesis that an augmented linear model whose design matrix is expanded to include weighted covariates is needed. This equivalence makes it very convenient to run the test in standard software. Fuller (1984) gives an approximate $\mathcal{F}$-test for the case of cluster samples within strata. The DuMouchel–Duncan test is extended to generalized linear models by Nordberg (1989). Pfeffermann (1993) extends comparison of weighted and unweighted parameter estimators to general likelihood-based problems with explicit estimators, and provides a Wald-type test statistic. Pfeffermann & Sverchkov (2003) extend the test to generalized linear models in which estimators are defined as the solutions to estimating equations.

Our approach builds on the idea of this latter class of tests comparing weighted and unweighted estimators. However, instead of making parametric assumptions, we propose a nonparametric test of informative selection.

Our nonparametric procedure is motivated by Gretton et al. (2012), who establish a kernel method for the classical two-sample-problem. In order to test if two samples are from different populations, they propose the Maximum Mean Discrepancy (MMD) test statistic, in which they find a smooth function f that is large on the points sampled from one population and small on the points sampled from the other population. The MMD test statistic is then the estimated difference between the expected function values on the two samples. By restricting the class of such smooth functions $\mathscr{F}$ to be the unit ball in a universal reproducing kernel Hilbert space (RKHS) $\mathscr{H}$ (Aronszajn (1950)), with reproducing kernel k, $\mathscr{F}_k = \{f : \|f\|_{\mathscr{H}} \leq 1\}$, Gretton et al. (2012) propose a computationally-feasible MMD statistic and derive its asymptotic distribution under the null and alternative hypotheses. Advantages of using $\mathscr{F}_k$ are summarized in Sriperumbudur et al. (2010).

We propose a novel nonparametric test for informative selection that uses the maximum mean discrepancy between the weighted sample and the unweighted sample, and establish its asymptotic distribution under the null hypothesis of noninformative selection. The test relies only on sample observations (possibly vector-valued) and sample weights. We show via simulation that the asymptotic test has correct size in finite samples and good power for a variety of informative alternatives.

This chapter is organized as follows. Notation and assumptions are introduced in Section 3.2, following the framework of Gretton et al. (2012). We prove the asymptotic result for our statistic in Section 3.3. In Section 3.4, we explore the power and size properties of our tests and compare to existing parametric tests and the cdf-based nonparametric tests introduced in Section 2, under different types and amounts of informativeness. In Section 3.5, we apply our methodology to data from a recreational fisheries survey, the Marine Recreational Information Program (MRIP). The proofs are deferred to the Appendix.

3.2 Notation and Assumptions

Consider a finite population $U = \{1, \ldots, N\}$. $\{y_i\}_{i \in U}$ are realization of $\{Y_i\}_{i \in U}$, independent and identically distributed (iid) with probability measure p. Only a subset of U is observed, and such subset s is called a sample. We use sample membership indicator ξ_i to denote if an element is being selected, $\xi_i = 1$ if $i \in s$, and $\xi_i = 0$ otherwise. The inclusion probability $\pi_i = \mathbb{E}\left[\xi_i \middle| Y_i = y_i\right]$. The sample weights w_i's are the inverse of the π_i's, $w_i = \pi_i^{-1}$. Sample size $n = |s|$.

Let p and q be probability measures and let Y and Y' be random variables defined on $\mathscr{Y}$. For a set of functions $\mathscr{F}$, the *maximum mean discrepancy (MMD)* between p and q over $\mathscr{F}$ is defined as

$$\mathrm{MMD}[\mathscr{F}, p, q] := \sup_{f \in \mathscr{F}} \left(\mathbb{E}_{y \sim p}\left[f(y)\right] - \mathbb{E}_{y' \sim q}\left[f(y')\right]\right). \tag{3.2.1}$$

The idea behind this metric is that we pick f to be large on p and small on q, so that MMD would be large if the two samples are from different distributions, while MMD is zero if $p = q$.

To sidestep the problem of choosing f, it is convenient to choose $\mathscr{F}$ to be the unit ball $\{\mathscr{F} : \|f\|_{\mathscr{H}} \leq 1\}$ in a universal reproducing kernel Hilbert space (RKHS) $\mathscr{H}$, in the sense of section 2.2 of Gretton et al. (2012), with universal kernel $k(\cdot, \cdot)$. With restriction that $\mathscr{Y}$ is compact, a universal RKHS is dense in $C(\mathscr{Y})$ with respect to the L^∞ norm. Steinwart (2002) shows that Gaussian and Laplace kernels are universal. Assume k is continuous, symmetric, positive definite and square integrable. Using RKHS properties and squaring MMD for convenience, Gretton et al. (2012) show that

$$\begin{aligned}
\mathrm{MMD}^2[\mathscr{F}, p, q] &= \left\{ \sup_{f \in \mathscr{H} : \|f\| \leq 1} \left(\mathbb{E}_{y \sim p}\left[f(y)\right] - \mathbb{E}_{y' \sim q}\left[f(y')\right]\right) \right\}^2 \\
&= \|\mathbb{E}_{y \sim p}\left[k(y, \cdot)\right] - \mathbb{E}_{y' \sim q}\left[k(y', \cdot)\right]\|_{\mathscr{H}}^2 && (3.2.2) \\
&= \mathbb{E}_{y \sim p, y' \sim p}\langle k(y, \cdot), k(y', \cdot)\rangle - 2\mathbb{E}_{y \sim p, y' \sim q}\langle k(y, \cdot), k(y', \cdot)\rangle \\
&\qquad + \mathbb{E}_{y \sim q, y' \sim q}\langle k(y, \cdot), k(y', \cdot)\rangle \\
&= \mathbb{E}_{y \sim p, y' \sim p}[k(y, y')] - 2\mathbb{E}_{y \sim p, y' \sim q}[k(y, y')] + \mathbb{E}_{y \sim q, y' \sim q}[k(y, y')], && (3.2.3)
\end{aligned}$$

where y and y' are independent random variables in each expectation.

Given samples $\{y_1, y_2, \ldots, y_n\}$ iid from p and $\{y'_1, y'_2, \ldots, y'_n\}$ iid from q, an unbiased and consistent estimator of $\mathrm{MMD}^2[\mathscr{F}, p, q]$ is then, from (3.2.3),

$$\mathrm{MMD}^2 = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{i \neq j}^{n} \left\{k(y_i, y_j) - 2k(y_i, y'_j) + k(y'_i, y'_j)\right\}.$$

Under the null hypothesis $p = q$, this statistic has mean zero, and Gretton et al. (2012) show that $n\mathrm{MMD}^2$ converges in distribution to an infinite linear combination of χ_1^2 random variables, with coefficients given by the eigenvalues of the centered kernel

$$h(y_i, y_j) \coloneqq k(y_i, y_j) - \mathbb{E}_y k(y, y_j) - \mathbb{E}_y k(y_i, y) + \mathbb{E}_{y,y'} k(y, y'), \tag{3.2.4}$$

where the eigen-decomposition is with respect to the common (null) probability measure $p = q$.

3.3 Methods

Our test statistic follows similar reasoning to the MMD of Gretton et al. (2012), but instead of comparing two independent samples from possibly different probability measures, it compares weighted and unweighted versions of the same sample. Under the alternative of informative selection, the superpopulation probability measure q is correctly estimated by using the survey weights w_i, which are inverse inclusion probabilities. By contrast, the probability measure p, estimated by the unweighted sample, is not q but

$$p(A) = \frac{\int_{y \in A} \pi(y)\, dq(y)}{\int_{y \in \mathcal{Y}} \pi(y)\, dq(y)}$$

for measurable sets $A \subset \mathcal{Y}$, where $\mathcal{Y}$ is the entire outcome space. That is, under the alternative the inclusion probability $\pi(y) = \mathbb{P}(\xi_i = 1 \mid y)$ depends on y and the design is informative. If $\pi(y) = \mathbb{P}(\xi_i = 1 \mid y) = \pi$, a constant independent of y, then the design is noninformative and $p(A) = q(A)$ for all measurable sets A.

Because we do not have independent samples, we construct a plug-in estimator of the population $\mathrm{MMD}^2[\mathscr{F}, p, q]$ not from (3.2.3) but from (3.2.2), then correct for bias. First, define

$$
\begin{aligned}
A &= \left\| \frac{1}{n}\sum_{i\in s} k(\cdot, y_i) - \frac{1}{n}\sum_{i\in s} \frac{w_i}{\overline{w}_s} k(\cdot, y_i) \right\|^2_{\mathscr{H}} \\
&= \left\| \frac{1}{n}\sum_{i\in s} \left(1 - \frac{w_i}{\overline{w}_s}\right) k(\cdot, y_i) \right\|^2_{\mathscr{H}} \\
&= \frac{1}{n^2}\sum_{i\in s}\sum_{j\in s} \left(1 - \frac{w_i}{\overline{w}_s}\right)\left(1 - \frac{w_j}{\overline{w}_s}\right) \langle k(\cdot, y_i) k(\cdot, y_j)\rangle_{\mathscr{H}} \\
&= \frac{1}{n^2}\sum_{i\in s}\sum_{j\in s} \left(1 - \frac{w_i}{\overline{w}_s}\right)\left(1 - \frac{w_j}{\overline{w}_s}\right) k(y_i, y_j),
\end{aligned}
$$

by properties of the inner product and the kernel in the RKHS. Since $\sum_{i\in s}(1 - w_i/\overline{w}_s) = 0$, we can replace k by h from (3.2.4) without changing the value:

$$
A = \frac{1}{n^2}\sum_{i\in s}\sum_{j\in s} \left(1 - \frac{w_i}{\overline{w}_s}\right)\left(1 - \frac{w_j}{\overline{w}_s}\right) h(y_i, y_j). \tag{3.3.1}
$$

Because (3.3.1) is not unbiased for $\mathrm{MMD}^2[\mathscr{F}, p, q]$, we remove the main diagonal. Further, we replace the theoretically-centered kernel h by its empirical estimator $\widehat{h}$ and normalize by n to obtain the following test statistic:

$$
n\mathrm{MMD}^2_{\widehat{h}} = \frac{1}{n-1}\sum\sum_{i,j\in s, i\neq j} \left(1 - \frac{w_i}{\overline{w}_s}\right)\left(1 - \frac{w_j}{\overline{w}_s}\right) \widehat{h}(y_i, y_j), \tag{3.3.2}
$$

where $\overline{w}_s = \sum_{i\in s} w_i/n$ and

$$
\widehat{h}(y_i, y_j) = k(y_i, y_j) - \frac{1}{n}\sum_{j\in s} k(y_i, y_j) - \frac{1}{n}\sum_{i\in s} k(y_i, y_j) + \frac{1}{n^2}\sum_{i\in s}\sum_{j\in s} k(y_i, y_j). \tag{3.3.3}
$$

We then have the following asymptotic result.

Theorem 3.3.1. *Assume that the design weights $\{w_i\}_{i\in s}$ are iid with mean μ_w and finite variance σ^2_w, and assume the null hypothesis that the weights are independent of sampled responses $\{y_i\}_{i\in s}$ that are iid q. Assume further that $\mathbb{E}_q\left[k^2(y, y')\right] < \infty$, where y, y' are iid q. Then, as $n \to \infty$,*

$$n\mathrm{MMD}^2_{\widehat{h}} \xrightarrow{\mathcal{L}} \sum_{l=1}^{\infty} \lambda_l \left(z_l^2 - \frac{\sigma_w^2}{\mu_w^2} \right), \tag{3.3.4}$$

where $\{z_l\}$ are iid $\mathcal{N}(0, \sigma_w^2/\mu_w^2)$, and λ_l are solutions to the eigenvalue equation

$$\int_{\mathcal{Y}} h(y, y')\psi_l(y)\, dq(y) = \lambda_l \psi_l(y').$$

Practical use of this theorem requires estimates of the eigenvalues, provided in the following result.

Remark 3.3.1.1. *Gretton et al. (2009) show that empirical estimates of the first n eigenvalues can be obtained as*

$$\widehat{\lambda}_l = \frac{1}{n}\nu_l,$$

where ν_l are the eigenvalues of the centered Gram matrix

$$\widetilde{K} := CKC,$$

$K := [k(y_i, y_j)]_{i,j \in s}$*, $C = I - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}}$ is a centering matrix, and $\mathbf{1}$ is an $n \times 1$ vector of 1's.*

The full implementation of the MMD hypothesis test then proceeds as follows. First, choose a kernel k (standard choices would be Gaussian or Laplacian densities). Next, determine the kernel bandwidth. An empirical rule that seems to work well is to use the median of the interpoint distances among the sample $\{y_i\}_{i \in s}$; for example, set the standard deviation for a Gaussian kernel or the scale parameter for a Laplacian kernel equal to this median value. Compute the kernel values $K = [k(y_i, y_j)]$ and the MMD statistic (3.3.2). Use the empirical mean and variance of the weights to estimate μ_w and σ_w^2, and use Remark 3.3.1.1 to estimate the first n eigenvalues $\{\lambda_l\}_{l=1}^n$. With these estimated values, simulate a large number of realizations from the limiting distribution (3.3.4) via

$$\sum_{l=1}^{n} \widehat{\lambda}_l \left(z_l^2 - \frac{\widehat{\sigma}_w^2}{\widehat{\mu}_w^2} \right),$$

where $\{z_l\}$ are iid $\mathcal{N}(0, \widehat{\sigma}_w^2/\widehat{\mu}_w^2)$. The p-value of the test is then the proportion of simulated realizations that are greater than the computed MMD statistics.

3.4 Simulation

3.4.1 Gestational Age Example

As described in Bonnéry et al. (2018), this simulation is motivated by the actual design of the 1988 National Maternal and Infant Health Survey (NMIHS), which oversampled low birthweight infants (Korn & Graubard (1999)). We simulate an unstratified design with $N = 15000$ and $n = 300$ instead of the stratified sampling design in the original setting. We assume Y = gestational age to be iid $\mathcal{N}(\theta, \sigma^2)$ in the superpopulation. Because birthweight and gestational age are highly correlated, oversampling lower birthweight infants means that inclusion probabilities and sampling weights depend on gestational age. Thus the design is informative.

We generate the inclusion probabilities by

$$\ln \pi \mid (Y = y) \sim \mathcal{N}(-\delta_0 - \delta y, \tau^2)$$

with

$$\delta_0 = -\ln \frac{n}{N} + \frac{\tau^2}{2} - \delta\theta + \frac{\delta^2\sigma^2}{2}.$$

We set $\theta = 39.853$, $\sigma^2 = 16.723$, $\tau^2 = 0.087$ and let δ range from 0 to 0.03 with grid size of 0.002. Lower values of δ means less informativeness, with noninformativeness when $\delta = 0$.

For each of the 1000 independent replicates, a Poisson sample of expected size $n = 300$ is selected. MMD is computed as in (3.3.2). Figure 3.1 shows the power of our method compared to DuMouchel and Duncan (DD, DuMouchel & Duncan (1983)), Pfeffermann (PFE, Pfeffermann (1993)), Kolmogorov–Smirnov (KS), and Cramér-von Mises (CvM) versus δ; informativeness increases with δ. All methods maintain approximately the correct size at the null hypothesis of noninformative selection, $\delta = 0$. In this example, MMD statistic shows competitive power relative to DD and PFE. The DD statistic shows the highest power while CvM power curve is slightly below DD. The KS statistic shows lowest power. PFE and MMD power curves lie between CvM and KS, and are almost right on top of each other. The reason that PFE does not perform as well as DD is that the PFE statistic looks for informativeness in both the mean (which is altered by informative selection in this example) and variance (which is not altered), while DD has the advantage of looking for informativeness only in the mean. In other simulations (not shown), the PFE shows approximately the same power as DD if we plug in the true variance in the likelihood. Here DD assumes the correct model and PFE uses the correct likelihood, while MMD, KS and CvM use only sample responses and their design weights. Essentially, our method achieves good power without any of the modeling cost of the parametric methods.

3.4.2 Two-Stage Gestational Age Example

We investigate the application of MMD to two-stage sampling via simulation, by extending the gestational age example of Section 3.4.1. Consider a finite population with $N_{PSU} = 15000$ primary sampling units (PSUs), each made up of $m = 10$ secondary sampling units (SSUs). We set Z_{ij} iid $\mathcal{N}(\theta/m, \sigma^2/m)$ for $i = 1, 2, \ldots, N_{PSU}$ and $j = 1, 2, \ldots, m$, so that the PSU totals $Y_i = \sum_{j=1}^{m} Z_{ij}$ are iid $\mathcal{N}(\theta, \sigma^2)$ across PSUs. We generate the inclusion probabilities of the first stage by

$$\ln \pi \mid Y = y \sim \mathcal{N}(-\delta_0 - \delta y, \tau^2)$$

with

$$\delta_0 = -\ln \frac{n}{N} + \frac{\tau^2}{2} - \delta\theta + \frac{\delta^2 \sigma^2}{2}.$$

We set $\theta = 39.853$, $\sigma^2 = 16.723$, $\tau^2 = 0.087$ and δ ranges from 0 to 0.03 with grid size of 0.002 as before. For each of 1000 independent replicates, a Poisson sample of the PSUs with expected sample size $n = 300$ is selected, and then a simple random sample without replacement (SRSWOR) within each selected PSU is selected, with sample size n_m.

The goal in this problem is inference for the PSU totals. The estimated PSU totals, $\widehat{y}_i = m \sum_{j \in s_i} z_{ij}/n_m$, are computed from the simple random sample s_i we collected in the second stage.

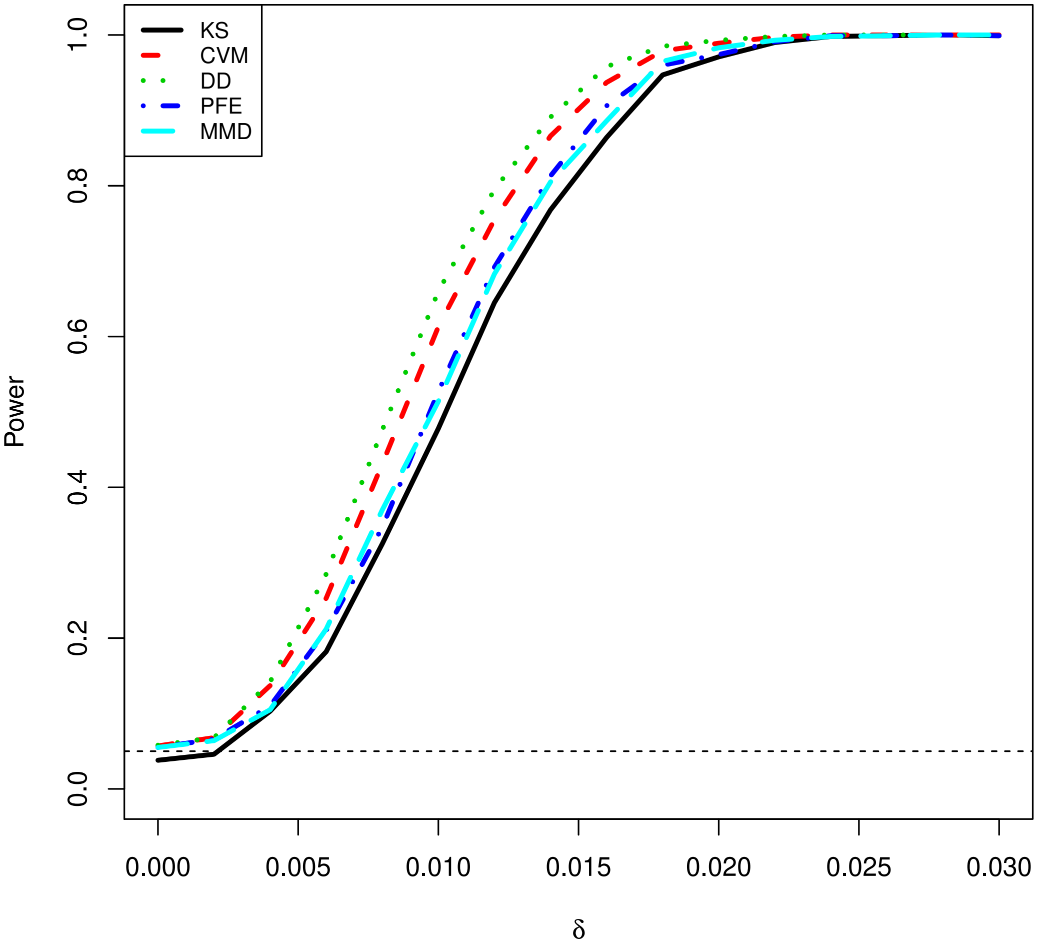

Figure 3.1: Power versus informativeness for DuMouchel–Duncan (DD; top curve), Pfeffermann (PFE; second lowest curve, dotdash), Kolmogorov–Smirnov (KS; bottom curve), Cramér–von Mises (CvM; second highest curve) and Maximum Mean Discrepancy (MMD; second lowest curve, longdash) tests, based on 1000 replicate Poisson samples of expected size $n = 300$ from the simulated gestational age population. Nominal size of all tests is $\alpha = 0.05$ (horizontal reference line). Informativeness increases with the value of δ, with $\delta = 0$ the noninformative null.

Then the calculations follow as in Section 3.4.1 by using estimated PSU totals $\widehat{y}_i$ in place of y_i and using PSU weights w_i. Figure 3.2 shows the power curves of all the statistics for various amount of informativeness with $n_m = 2$ and $n_m = 6$. The single stage sampling in Section 3.4.1 can be treated as a special case of two-stage sampling with $n_m = 10$. In spite of the subsampling, the tests maintain the correct test size and the same relative power ordering as in Figure 3.1, though power decreases as the subsample size n_m decreases.

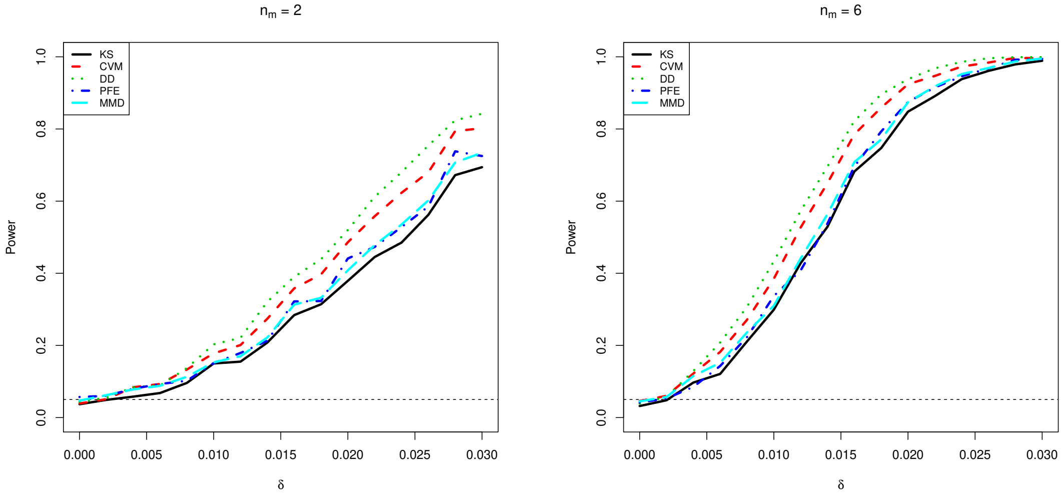

Figure 3.2: Power versus informativeness for DuMouchel–Duncan (DD; top curve), Pfeffermann (PFE; second lowest curve, dotdash), Kolmogorov–Smirnov (KS; bottom curve), Cramér–von Mises (CvM; second highest curve) and Maximum Mean Discrepancy (MMD; second lowest curve, longdash) tests, based on 1000 replicate two-stage samples, with Poisson samples of expected size $n = 300$ PSUs in stage one, and simple random samples without replacement of size $n_m = 2$ (left panel) and $n_m = 6$ (right panel) in stage two. Nominal size of all tests is $\alpha = 0.05$ (horizontal reference line). Informativeness increases with the value of δ, with $\delta = 0$ the noninformative null.

3.4.3 Multidimensional Gestational Age Example

One big advantage of our MMD approach is that vector-valued responses can be tested simultaneously, exactly as with a scalar response. Instead of building parametric models for each scalar response, or running a nonparametric test for each scalar response, we can run a single test to detect if there is informative selection anywhere among the responses.

To illustrate, we generalize the gestational age example in Section 3.4.1 to the m-dimensional case. The response vector of interest, $\mathbf{Y} = (Y_1, Y_2, \dots, Y_m)^\intercal$, is normally distributed in the superpopulation with mean vector $\boldsymbol{\theta} = (\theta, \theta, \dots, \theta)^\intercal$ and covariance matrix $\boldsymbol{\Sigma} = \text{diag}(\sigma^2, \sigma^2, \dots, \sigma^2)$. The finite population vector responses, $\mathbf{y}_i = (y_{1i}, y_{2i}, \dots, y_{mi})^\intercal$, are iid realizations from the superpopulation model for $i = 1, 2, \dots, N$.

The inclusion probabilities are generated by an informative mechanism that depends only on the first component of the response vector:

$$\ln \pi \mid y_1 \sim \mathcal{N}(-\delta_0 - \delta y_1, \tau^2)$$

with

$$\delta_0 = -\ln \frac{n}{N} + \frac{\tau^2}{2} - \delta\theta + \frac{\delta^2\sigma^2}{2}.$$

We set $\theta = 39.853$, $\sigma^2 = 16.723$, $\tau^2 = 0.087$ and let δ range from 0 to 0.03 with grid size of 0.002. Lower values of δ correspond to less informativeness, with noninformativeness when $\delta = 0$. Four different dimensions are considered ($m = 1, 4, 16, 64$), so that the true informative mechanism is increasingly obscured as the dimension increases.

For each of the 1000 independent replicates, we draw a Poisson sample of expected size $n = 300$. The MMD statistic is calculated as in (3.3.2). Figure 3.3 shows the power of the MMD test versus different amounts of informativeness for the four dimensions. No other tests are compared, as we are not aware of other tests for informative selection in the multi-dimensional case.

The scalar case, $m = 1$, is the same as in Section 3.4.1. As dimension grows (but expected sample size remains fixed), the power decreases and it is harder to detect the informativeness in the design. The MMD test has no guidance about which, if any, of the m dimensions might be informative. As expected, power decreases as dimension increases, but even in the highest dimension considered, the MMD test has considerable power against higher levels of informativeness. All cases show approximately the correct test size at $\delta = 0$.

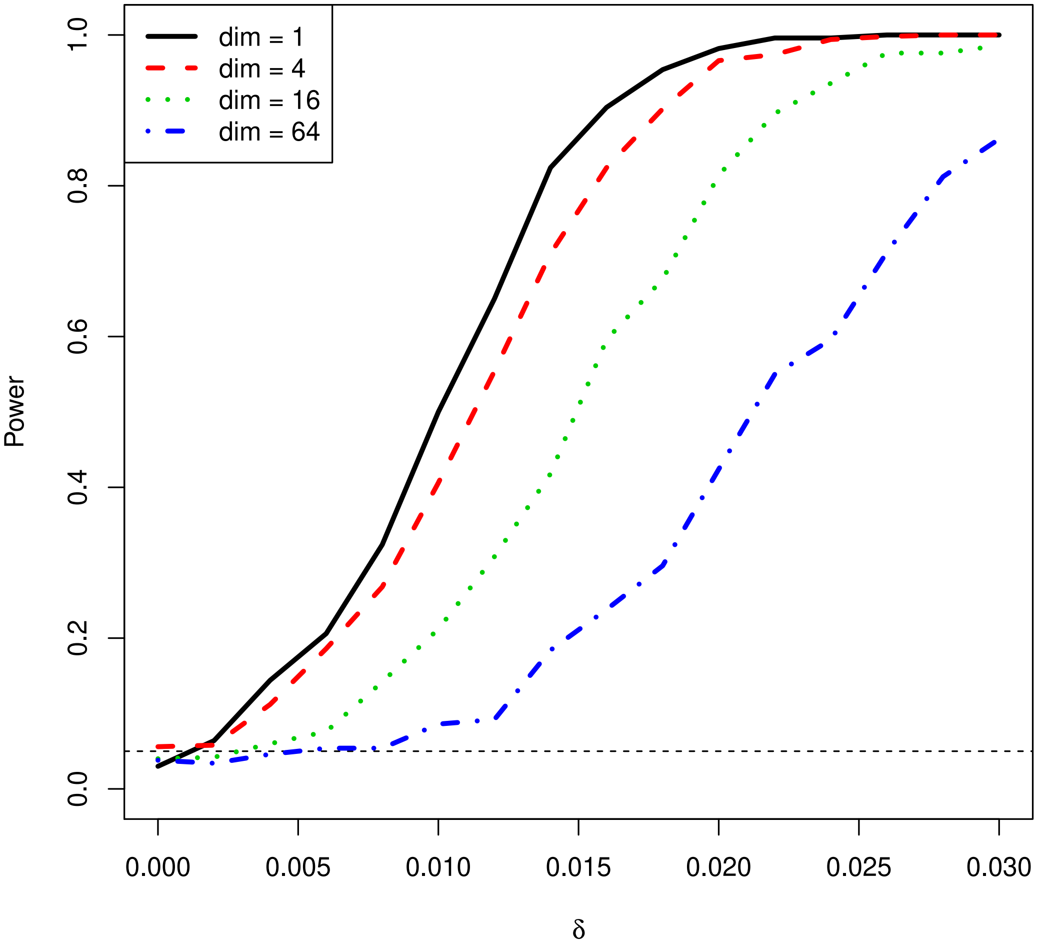

Figure 3.3: Power versus informativeness for Maximum Mean Discrepancy (MMD) test with various response dimension (dim = 1; top curve), (dim = 4; second highest curve), (dim = 16; second lowest curve) and (dim = 64; bottom curve), based on 1000 replicate Poisson samples of expected size $n = 300$ from the simulated gestational age population. Nominal size of all tests is $\alpha = 0.05$ (horizontal reference line). Informativeness increases with the value of δ, with $\delta = 0$ the noninformative null.

3.4.4 Scaled *t* Distribution

As the informativeness exists in the mean for the gestational age example, both DD and PFE show good power as expected: they assume the correctly-specified mean model and likelihood. In the following t example, the informativeness exists in the variance.

Our variable of interest y_i is generated as follows:

$$y_i = \mu + \sigma \frac{z_i}{\sqrt{v_i/\nu}} \sqrt{\frac{\nu - 2}{\nu}} = \mu + \sigma_i z_i,$$

where $\{z_i\}$ are iid $\mathcal{N}(0, 1)$, independent of $\{v_i\}$ iid χ^2_ν. The error terms here are distributed as scaled t_ν, with mean 0 and variance σ^2 for $\nu > 2$, and

$$y_i \mid v_i \sim \mathcal{N}(0, \sigma_i^2).$$

Let $\{v_i^*\}$ be iid χ^2_ν, independent of $\{z_i\}$ and $\{v_i\}$, and set

$$\tau_i = \sigma \frac{1}{\sqrt{v_i^*/\nu}} \sqrt{\frac{\nu - 2}{\nu}},$$

which has the same distribution as σ_i but is not used in generating y_i. Define $d_i = \rho\sigma_i + (1 - \rho)\tau_i$ and set the inclusion probabilities as

$$\pi_i = n d_i \left(\sum_{i \in U} d_i \right)^{-1},$$

where $\rho \in [0, 1]$ is a constant to control the amount of informativeness in the design. Samples are selected by Poisson sampling with inclusion probability proportional to size d_i. The motivation for this example is that designs with $\pi_i \propto \sigma_i$ minimize the unconditional variance, with respect to model and design, of the Horvitz-Thompson estimator of the y-total, and d_i is a proxy for σ_i.

We set $\mu = 39.853$, $\sigma^2 = 6.123$, $\nu = 5$, and let ρ range from the noninformative null at $\rho = 0$ to the highly-informative, optimal design at $\rho = 1$ with grid size of 0.025. For each of 1000

independent replicates, a Poisson sample of expected size $n = 300$ is selected. We compute MMD statistic as in (3.3.2).

Figure 3.4 shows the power of our method compared to DD, PFE, KS and CvM versus ρ; informativeness increases with ρ. All tests have approximately the correct test size at the noninformative null, $\rho = 0$. As ρ increases, DD has a very low amount of power, because the weighted and unweighted estimates of the mean will differ significantly, by chance, due to large variation, even though both the weighted and unweighted estimates are actually unbiased for the mean. That is, the DD test sometimes “guesses the right answer” when it rejects the null by mistakenly attributing the difference to bias caused by informative selection. PFE uses the correctly-specified likelihood and has the most power. KS and CvM tests are much less powerful than PFE. MMD, however, shows good power that is quite comparable to PFE for a wide range of ρ. Since MMD uses only the sample observations and their weights, and requires no parametric modeling at all, it is always worth trying as a test for informative selection.

3.4.5 Multidimensional Scaled *t* Distribution

MMD shows good power in multidimensional testing of informativeness in the mean in Section 3.4.3. Here we investigate its power in multidimensional testing of informativeness in the variance. We generalize the scaled t example in Section 3.4.4 to the m-dimensional case. The response vector of interest is $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_m)^\intercal$. The finite population vector responses, $\mathbf{y}_i = (y_{1i}, y_{2i}, \ldots, y_{ji}, \ldots, y_{mi})^\intercal$, are iid realizations from the following model for $i = 1, 2, \ldots, N$ and $j = 1, 2, \ldots, m$:

$$y_{ji} = \mu + \sigma \frac{z_i}{\sqrt{v_{ji}/\nu}} \sqrt{\frac{\nu - 2}{\nu}} = \mu + \sigma_{ji} z_i,$$

where $\{z_i\}$ are iid $\mathcal{N}(0, 1)$, independent of $\{v_{ji}\}$ iid χ^2_ν. The error terms here are distributed as scaled t_ν, with mean 0 and variance σ^2 for $\nu > 2$, and

$$y_{ji} \mid v_{ji} \sim \mathcal{N}(0, \sigma^2_{ji}).$$

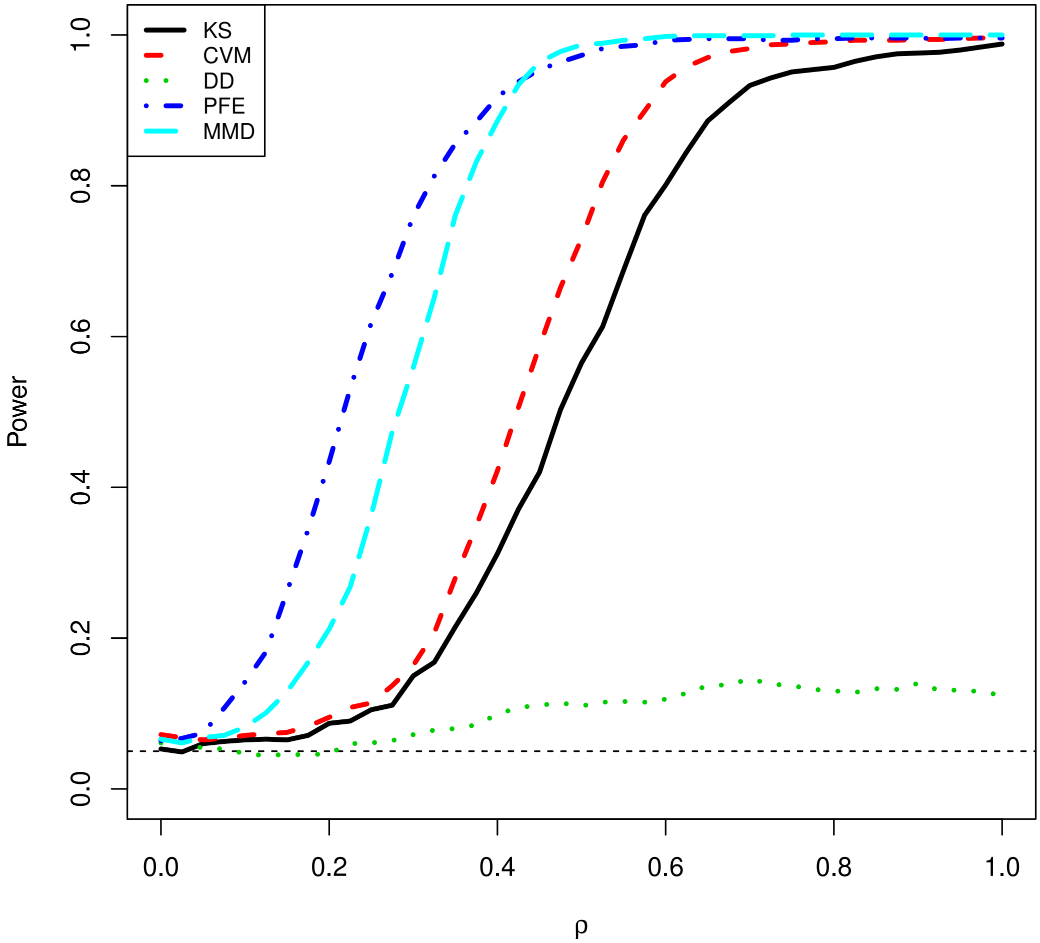

Figure 3.4: Power versus informativeness for DuMouchel–Duncan (DD; bottom curve), Pfeffermann (PFE; top curve), Kolmogorov–Smirnov (KS; second-lowest curve), Cramér–von Mises (CvM; third highest curve) and Maximum Mean Discrepancy (MMD; second highest curve) tests, based on 1000 replicate Poisson samples of expected size $n = 300$ from the simulated scaled-t population. Nominal size of all tests is $\alpha = 0.05$ (horizontal reference line). Informativeness increases with the value of ρ, with $\rho = 0$ the noninformative null.

Let $\{v_i^*\}$ be iid χ^2_ν, independent of $\{z_i\}$ and $\{v_i\}$, and set

$$\tau_i = \sigma \frac{1}{\sqrt{v_i^*/\nu}} \sqrt{\frac{\nu - 2}{\nu}},$$

which has the same distribution as σ_{ji} but is not used in generating $\mathbf{y}_i$. Define $d_i = \rho\sigma_{1i} + (1-\rho)\tau_i$ and set the inclusion probabilities as

$$\pi_i = nd_i \left(\sum_{i \in U} d_i \right)^{-1},$$

where $\rho \in [0, 1]$ is a constant to control the informativeness of the design. Here d_i only depends on the first component of $\boldsymbol{\sigma}_i$, thus only depends on the first component of $\mathbf{y}_i$. As π_i is proportional to d_i, π_i only depends on the first component of $\mathbf{y}_i$. Four different dimensions are considered ($m = 1, 4, 16, 64$), so that the informative mechanism is getting obscured as the dimension increases.

We set $\mu = 39.853$, $\sigma^2 = 6.123$, $\nu = 5$, and let ρ increase from the null of noninformativeness at $\rho = 0$ to the highly-informative, optimal design (for the first component of the response vector) at $\rho = 1$, with grid size of 0.025. For each of 1000 independent replicates, a Poisson sample of expected size $n = 300$ is selected. The MMD statistic is computed as in (3.3.2). Figure 3.5 shows the power of the MMD test versus various amounts of informativeness for the four dimensions.

The scalar case, $m = 1$, is the same as in Section 3.4.4. As expected, power decreases as dimension increases. The power of MMD in $m = 4$ dimensions is similar to the power of KS in only $m = 1$ dimension. MMD does not show as good power as it did in Section 3.4.3 for high dimensions. This reflects the difficulty in detecting informativeness in the variance. However, the MMD test still has some power against higher levels of informativeness, and all cases show approximately the correct test size at $\rho = 0$.

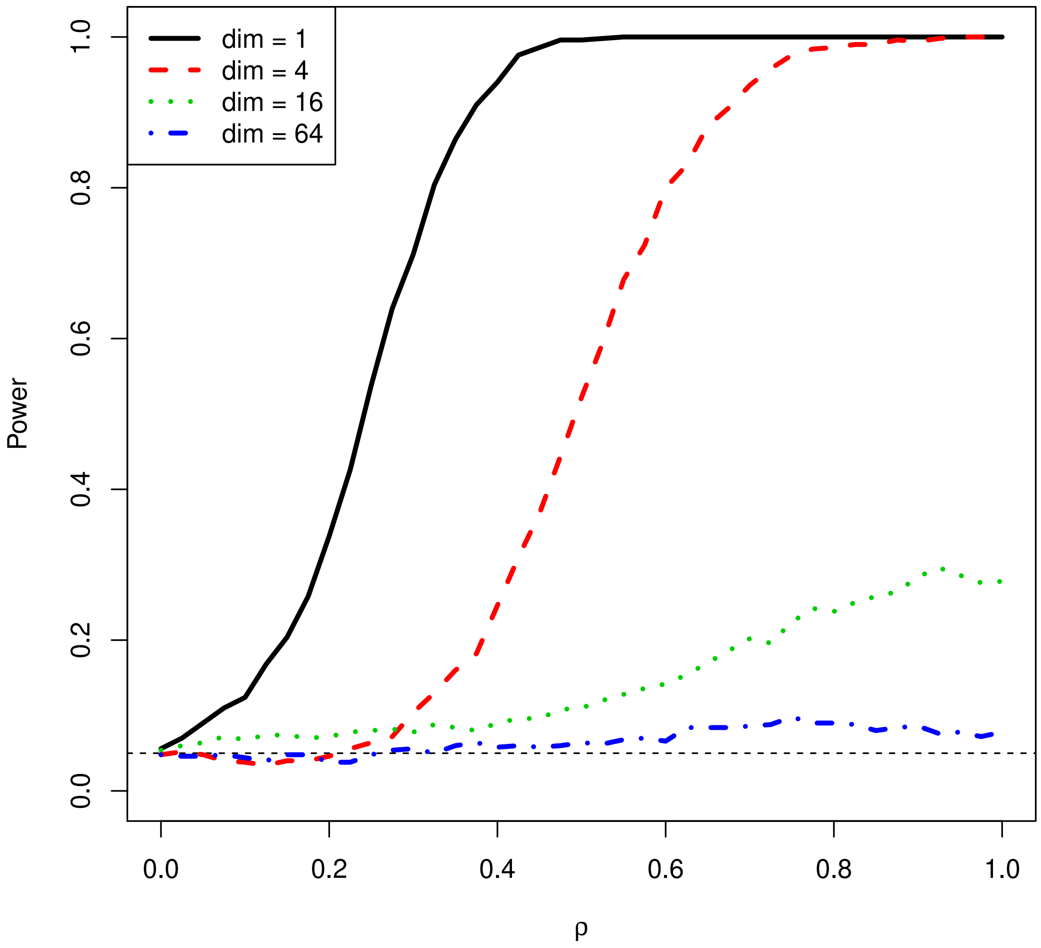

Figure 3.5: Power versus informativeness for Maximum Mean Discrepancy (MMD) test with various response dimension (dim = 1; top curve), (dim = 4; second highest curve), (dim = 16; second lowest curve) and (dim = 64; bottom curve), based on 1000 replicate Poisson samples of expected size $n = 300$ from the simulated scaled-t population. Nominal size of all tests is $\alpha = 0.05$ (horizontal reference line). Informativeness increases with the value of ρ, with $\rho = 0$ the noninformative null.

3.4.6 Pretest Estimators for Scaled t Distribution

As mentioned in Section 3.1, if the sampling design is noninformative, we should use the unweighted estimator for efficiency. Otherwise, we should adopt design-weighted estimators or other estimation methods that incorporate design information.

Our test suggests an alternative, "pretest estimator": use the unweighted estimator if the test fails to reject the null hypothesis of noninformative selection, or use the weighted estimator if the test rejects the null.

We compare the root mean squared error (rMSE) of the weighted, unweighted and pretest estimators (using KS, CvM or MMD as the test) via simulation for the mean, median, upper quartile and 90th percentile of the scaled t distribution as in Section 3.4.4. Figure 3.6 shows the rMSE of these estimators as a function of the amount of informativeness. In each panel, the dotted line shows the rMSE of the unweighted estimator, which increases due to increasing variability of the sample as ρ increases, and due to bias under informative selection for the quantiles (but not the mean). The rMSE of the weighted estimator is represented by the dashed line sloping down from the left. The weighted estimator loses some efficiency due to unnecessary weighting when ρ is small, but reduces bias without an increasing price in variance as ρ increases. The other three curves, solid for KS and dash-dot for CvM, two-dash for MMD, show the rMSE of the pretest estimators. The pretest estimators have both advantages of efficiency in low informativeness and unbiasedness in high informativeness. The three pretest estimators behave similarly, each dominating the weighted estimator for low values of ρ, dominating the unweighted estimator for intermediate values of ρ, and converging to the rMSE of the weighted estimator for high values of ρ.

The upper left panel of Figure 3.6 shows the estimation for the mean. The weighted estimator does not lose much efficiency in low informativeness, and pretest estimators have large variance in moderate informativeness as the unweighted estimator has large variance for moderate to large informativeness. So the weighted estimator is the best one for estimating the mean. Estimation for the median is shown on the upper right. With the increase in ρ, the unweighted estimator has huge bias and thus large rMSE, while the weighted estimator loses some efficiency for lower ρ, but has fairly stable rMSE overall. Our three pretest estimators have both advantages of efficiency in low informativeness and unbiasedness in high informativeness. The three pretest estimators dominate the weighted estimator for low values of ρ, dominate the unweighted estimator for intermediate values of ρ, and converge to the rMSE of the weighted estimator for high values of ρ. Estimation of upper quartile and 90th percentile are shown in the lower left and lower right. All pretest estimators gain efficiency in low informativeness, have some bias for intermediate ρ, and produce similar rMSE overall as the weighted estimator, with MMD showing smaller rMSE than KS and CvM. In sum, our pretest estimators dominate the weighted and unweighted estimators of quantiles in this example, and might have better estimation properties than the weighted estimator for certain problems.

3.5 Application to a Recreational Angling Survey

We apply the nonparametric tests of informative selection to recreational angling data from the 2016 Marine Recreational Information Program (MRIP) in South Carolina. MRIP measures the number of fishing trips taken by recreational anglers in saltwater, along with the number of fish of each species caught by the anglers. Data on recreational angling are important for understanding fish stocks and sustainable management of fisheries.

Both shore fishing and boat fishing are of interest in MRIP. In this example, we focus on boat trips. To estimate characteristics of the population of all boat trips in South Carolina in 2016, MRIP uses two complementary surveys: an on-site "intercept" survey that collects catch by species information via angler interviews at the fishing site, and an off-site "effort" survey that collects information on the number of angler trips via self-administered questionnaire (mail-out/mail-back). The two sources of information are combined into the weights for the intercept survey, so we focus here on the details of that survey.

In MRIP, the intercept survey of boat trips is obtained by constructing a frame of publicly-accessible sites where boats can return to shore; crossing those sites with days in the fishing sea-

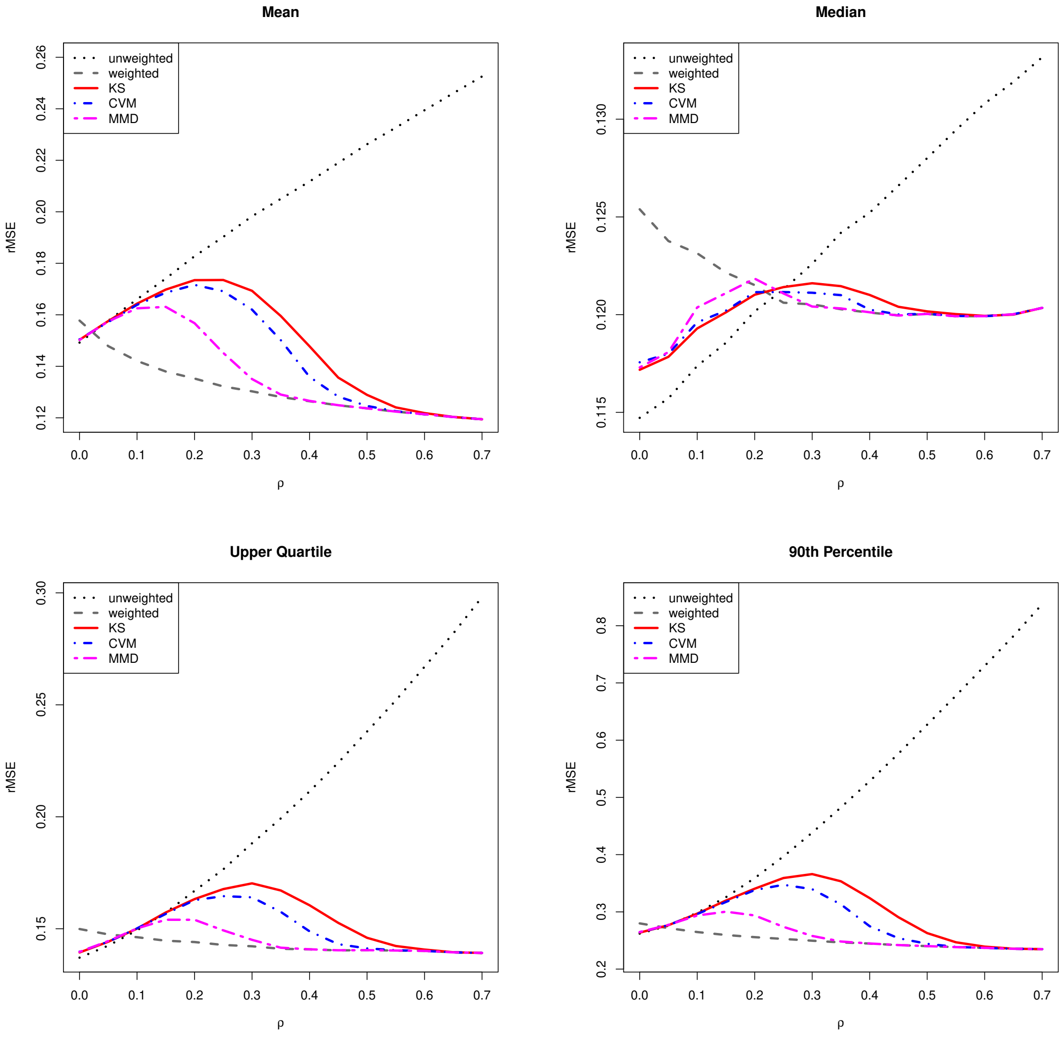

Figure 3.6: rMSE of unweighted estimator (increasing dotted curve), weighted estimator (decreasing dashed curve) and pretest estimators with KS statistic (solid curve), CvM statistic (dashed-dotted curve) and MMD statistic (two-dashed curve) for the mean (upper left panel), median (upper right panel), upper quartile (lower left panel) and 90th percentile (lower right panel). Results based on 1000 replicate Poisson samples of expected size $n = 300$ from the simulated scaled t population. Nominal size of tests for pretest estimators is $\alpha = 0.05$. Informativeness increases with the value of ρ, with $\rho = 0$ the noninformative null.

Table 3.1: p-values of Kolmogorov–Smirnov (KS), Cramér–von Mises (CvM), DuMouchel–Duncan (DD) and Maximum Mean Discrepancy (MMD) statistics for eight different response variables (number of anglers and catch for seven species) individually, and for combinations of all response variables, all but Anglers and all but Anglers and Black sea bass.

Variable of Interest	KS	CvM	DD	MMD
Anglers	0.020	0.007	0.027	0.022
Red drum	0.998	0.930	0.763	0.705
Black sea bass	0.095	0.091	0.060	0.029
Bluefish	0.650	0.569	0.863	0.286
Black drum	0.585	0.603	0.836	0.222
Wahoo	0.895	0.703	0.575	0.284
Gag grouper	0.815	0.655	0.300	0.229
Atlantic croaker	0.719	0.637	0.388	>0.999
All	–	–	–	0.032
All$\setminus$Anglers	–	–	–	0.082
All$\setminus\{$Anglers, Black sea bass$\}$	–	–	–	0.640

son to get “site-days”; stratifying site-days spatially (using contiguous South Carolina counties as strata) and temporally (using five two-month waves: March–April,..., November–December); obtaining a stratified sample of site-days; and intercepting all boat trips on selected site-days. The stratified sampling is conducted with probabilities proportional to estimated fishing activity for the site-days (“pressures”). The weights for MRIP reflect these unequal probabilities of selection and also reflect other adjustments, particularly from the effort survey.

Table 3.1 shows the p-values of KS and CvM statistics as described in Chapter 2, DD statistic as in DuMouchel & Duncan (1983) and MMD statistic for various variables. Anglers is significant for all the tests at test size $\alpha = 0.05$. Black sea bass shows significance by MMD at test size $\alpha = 0.05$ while being significant at test size $\alpha = 0.1$ for other tests. From the nature of the design, we know Anglers ought to be informative, as surveys were taken on the days and sites that were more likely to have anglers. Further, if catch for a species is correlated with number of anglers, we might expect the design to be informative for that species. After investigating the plots of Anglers with catch for the seven species, we do not see a strong correlation between them except for Black sea bass, with moderate correlation 0.332. Thus, it makes sense that the test finds evidence of informative selection for Black sea bass while failing to reject the noninformative null for the other species. MMD shows smaller p-value than KS, CvM and DD in detecting Black sea bass, and in fact is the only test that rejects the null at $\alpha = 0.05$.

Because our MMD test can also be applied to vector-valued responses, we have also provided in Table 3.1 (below the horizontal dashed line) the p-values for MMD applied to all response variables, all except Anglers, and all except Angler and Black sea bass. The MMD test rejects the null hypothesis of noninformative selection for all responses at $\alpha = 0.05$, and for all responses but Anglers at $\alpha = 0.1$. Once the two informative dimensions of Anglers and Black sea bass are removed from the vector, MMD fails to reject the noninformative null. These multivariate results are consistent with the individual results in the rest of the table.

3.6 Discussion

This chapter provides a nonparametric method of testing for informative selection by following the idea of one class of parametric tests that compare weighted and unweighted estimators of parameters. The test statistic is constructed by measuring the MMD between weighted and unweighted samples in RKHS. Power analysis shows the competitiveness of our method to other parametric methods. Moreover, the MMD statistic does not require parametric assumptions and can be widely applied to all situations. An important issue in the application of the MMD-based tests is the selection of the kernel width. A heuristic solution that we adopt is to pick the kernel width as the median distance between points in the sample. The optimum choice of kernel width is an ongoing area of research. Also, the choice of new kernels could lead to more powerful tests. The class of functions considered in this chapter is the RKHS, but this might be extended to other, more general, classes of functions. MMD measures the discrepancies in terms of norms of differences of mean embeddings. Other discrepancies, such as kernel Fisher discriminant and Kullback-Leibler divergence, could be investigated rather than the difference of RKHS means.

Chapter 4

A Small Area Estimation Approach for Reconciling Mode Differences in Two Surveys of Recreational Fishing Effort

4.1 Introduction

For decades, the National Marine Fisheries Service (NMFS) has conducted the Coastal Household Telephone Survey (CHTS) to collect recreational saltwater fishing effort (the number of fishing trips) from shore and private boat anglers in 17 US states along the coasts of the Atlantic Ocean and the Gulf of Mexico: Alabama, Connecticut, Delaware, Florida, Georgia, Louisiana, Maine, Maryland, Massachusetts, Mississippi, New Hampshire, New Jersey, New York, North Carolina, Rhode Island, South Carolina, and Virginia. Data collection occurs during a two-week period at the end of each two-month sample period (or "wave"), yielding six waves for each year. However, samples are not obtained for every wave in every state; for example, many states have no wave 1 sample, reflecting minimal fishing effort during January and February in those states.

The CHTS uses random digit dialing (RDD) for landlines of households in coastal counties. RDD suffers from several shortcomings in this context, such as the inefficiency at identifying anglers (National Research Council, 2006), the declining response rate for telephone surveys (Curtin et al., 2005), and the undercoverage of anglers due to the increase in wireless-only households (Blumberg & Luke, 2013). Thus, after some experimentation, NMFS implemented the new Fishing Effort Survey (FES) that involves mailing questionnaires to a probability sample of postal addresses (Andrews et al., 2014).

The telephone-based CHTS and the mail-based FES have obvious methodological differences. The two surveys have different coverage properties, because they use very different frames: RDD of landlines for CHTS versus address-based sampling, with oversampling of addresses matched to licensed anglers, for FES. They have different nonresponse patterns, with overall FES response rates nearly three times higher than CHTS response rates (Andrews et al., 2014). Finally, the measurement processes are fundamentally different, due to the differences in asking about angling activity over the phone versus a paper form.

Due at least in part to these methodological differences, there is a large discrepancy between the effort estimates from the CHTS and the FES estimates. Whatever the reasons for the discrepancy, it is of interest to fisheries managers and stock assessment scientists to be able to convert from the "units" of the telephone survey estimates to those of the mail survey estimates, and vice versa. This conversion is known as "calibration" in this context, and is not to be confused with the calibration method common in complex surveys. The calibration allows construction of a series of comparable estimates across time.

The data used for the calibration exercise come from the CHTS for most states and waves from 1982 to 2017, and from the FES for states and waves from 2015 to 2017. For each survey, the data consist of estimated total effort for shore fishing and for private boat fishing, along with estimated design variances and sample sizes, for each available state and wave.

The literature on reconciling estimates from more than one survey is sparse. J. Van Den Brakel et al. (2020) review different methods to measure discontinuities due to a survey process redesign. Depending on whether there is an overlapping period between the old and new surveys, how long such a period lasts, and how the old survey switches to the new survey, the problem can be divided into the following cases. For parallel data collection, where data is collected under the old and new designs alongside each other for a certain period, design-based methods in J. A. Van Den Brakel (2008, 2013), state-space models in J. A. Van Den Brakel (2008, 2010) and small area estimation models in Pfeffermann (2002, 2013) and Rao & Molina (2015) can be adopted, depending on the length of the parallel run and the sample sizes. For the phase-in approach, where changeover to the new design is done by a gradual roll-out, similar methods to those used for a parallel run can be implemented. For the case where there is no overlap at all, state-space models are recommended.

The methodology described here uses effort estimates transformed via natural logarithms, for either shore or private boat fishing. Let $\widehat{M}_{st}$ denote the estimated log-effort based on the mail survey in state s and year-wave t and let $\widehat{T}_{st}$ denote the estimated log-effort based on the telephone survey. We build a model that assumes that both mail and telephone estimates target a common underlying time series of true effort, but that each survey estimate is distorted both by sampling error and nonsampling error. The true effort series is further described with a classical time series model consisting of trend, seasonal, and irregular components. The sampling error series have properties that are well-understood based on features of the corresponding sampling designs, including well-estimated design variances. The nonsampling error cannot be completely disentangled from the true effort series. But given the overlap of mail and telephone estimates for some states and waves, the difference in the nonsampling errors can be estimated, and can be modeled with available covariates to allow extrapolation forward or backward in time. This extrapolation is a key part of the calibration procedure.

The combined model for the two sets of estimates and the underlying true effort series is a linear mixed model of a type that commonly appears in the context of area-level small area estimation, where it is known as the Fay-Herriot model (Fay & Herriot, 1979). In Fay-Herriot, it is standard to treat design variances as known. Our design variances are based on moderate to large sample sizes (minimum size $n = 39$) in each state and wave and so are well-estimated by the standards of small area estimation. A complication is that our design variances are on the original effort scale rather than the log scale. As an alternative to standard Taylor linearization, we develop a novel approach to transforming the estimated design variances that ensures analytic consistency between our mean model and our variance model.

The Fay-Herriot methodology leads to empirical best linear unbiased predictors (EBLUPs) of the mail target or the telephone target, and these constitute our calibrated effort series. Unlike the standard Fay-Herriot context, the EBLUPs require prediction at new sets of covariates. We adapt standard Mean Square Error (MSE) approximations and estimates to this non-standard situation, and evaluate their performance via simulation. Finally, we apply the methods to the problem of reconciling past telephone survey estimates to the mail survey.

4.2 Model

4.2.1 Mean Model

We fix attention on one type of fishing behavior, either shore or private boat: the model development is identical in both cases. We assume that the telephone effort estimate $\widehat{T}_{st}$ is a design-unbiased estimator of the "telephone target" T_{st}, which includes both the true effort and survey mode effects due to the telephone methodology, while the mail effort estimate $\widehat{M}_{st}$ is a design-unbiased estimator of the "mail target" M_{st}, which includes both the true effort and survey mode effects due to the mail methodology. That is,

$$\widehat{T}_{st} = T_{st} + e^T_{st} \text{ and } \widehat{M}_{st} = M_{st} + e^M_{st}$$

where the sampling errors $\{e^T_{st}\}$ and $\{e^M_{st}\}$ have zero mean under repeated sampling.

We assume that both the telephone target and the mail target contain the true effort series, which is further assumed to contain state-specific trends, due in part to changing state population sizes, state-specific seasonal effects that vary wave to wave, and irregular terms that are idiosyncratic effects not explained by regular trend or seasonal patterns. We model state-specific trends by using annual state-level population estimates from the US Census Bureau US Census Bureau (2016) on a log scale. We model a general seasonal pattern via indicators for the two-month waves, and allow the seasonal pattern to vary from state to state. The remaining irregular terms, denoted $\{\nu_{st}\}$ below, represent real variation not explained by the regular trend plus seasonal pattern, and are modeled as iid random variables with mean zero and unknown variance, ψ.

The survey mode effects present in the telephone and mail targets are nonsampling errors, including potential biases due to coverage error (population $\neq$ sampling frame), nonresponse error (sample $\neq$ respondents), and measurement error (true responses $\neq$ measured responses). These effects may have their own trend and seasonality: for example, due to changes in the quality of the frame over time, changes in response rates over years or waves, changes in implementation of measurement protocols over time, etc. These nonsampling errors thus cannot be completely disentangled from the true effort series (a problem in every survey).

Because of the availability of overlapping effort estimates, however, the difference in the effort estimates is an unbiased estimator of the difference in the survey mode effects. These differences can then be modeled and extrapolated to other time points that do not have overlapping data, allowing calibration from the telephone target to the mail target, and vice versa. The extrapolation requires a model and suitable covariates, which in this setting means covariates that explain the change in measurement error, nonresponse error, or coverage error over time. The calibration thus relies critically on extrapolation, with the usual caveat that the calibrated values may be badly wrong if the model does not hold over the full range of time.

The changing proportion of wireless-only households is a potential covariate for explaining changes in coverage error over time for the landline-only telephone survey. Accordingly, we obtained June and/or December wireless-only proportion estimates for each state from 2007–2015 from the National Health Interview Survey, conducted by the National Center for Health Statistics (Blumberg & Luke, 2013). We transformed these proportions via empirical logits and fitted the transformed values as state-specific lines with a slope change in 2010. The fitted model has an adjusted R^2 value of 0.9948. Transforming back to proportions and extrapolating backward in time yields a series $\{w_{st}\}$ that is approximately zero prior to the year 2000.

Either trend or seasonal could contain survey mode effects. Accordingly, we allow for the possibility that trend and seasonal are different for mail versus telephone, and in particular we allow for the possibility that either trend or seasonal can change with the level of wireless.

Our combined model then assumes

$$
\begin{aligned}
\widehat{T}_{st} &= T_{st} + e_{st}^{T} \\
T_{st} &= \boldsymbol{a}_{st}^{\intercal}\boldsymbol{\alpha} + 0 \cdot \boldsymbol{b}_{st}^{\intercal}\boldsymbol{\mu} + w_{st}\boldsymbol{c}_{st}^{\intercal}\boldsymbol{\gamma} + \nu_{st} \\
&= \left[\boldsymbol{a}_{st}^{\intercal}, \mathbf{0}^{\intercal}, w_{st}\boldsymbol{c}_{st}^{\intercal}\right]\boldsymbol{\beta} + \nu_{st} \\
&= \boldsymbol{x}_{Tst}^{\intercal}\boldsymbol{\beta} + \nu_{st} \\
\widehat{M}_{st} &= M_{st} + e_{st}^{M} \\
M_{st} &= \boldsymbol{a}_{st}^{\intercal}\boldsymbol{\alpha} + 1 \cdot \boldsymbol{b}_{st}^{\intercal}\boldsymbol{\mu} + 0 \cdot \boldsymbol{c}_{st}^{\intercal}\boldsymbol{\gamma} + \nu_{st} \\
&= = \left[\boldsymbol{a}_{st}^{\intercal}, \boldsymbol{b}_{st}^{\intercal}, \mathbf{0}^{\intercal}\right]\boldsymbol{\beta} + \nu_{st} \\
&= \boldsymbol{x}_{Mst}^{\intercal}\boldsymbol{\beta} + \nu_{st},
\end{aligned}
\tag{4.2.1}
$$

where

- $\boldsymbol{a}_{st}$ is a vector of known covariates, including intercept, log(population), state indicators, wave indicators, and state by log(population) and state by wave interactions;
- $\boldsymbol{b}_{st}$ and $\boldsymbol{c}_{st}$ are subvectors from $\boldsymbol{a}_{st}$;
- $\boldsymbol{\beta}^{\intercal} = [\boldsymbol{\alpha}^{\intercal}, \boldsymbol{\mu}^{\intercal}, \boldsymbol{\gamma}^{\intercal}]$ is a vector of unknown regression coefficients;
- the sampling errors $\{e_{st}^{T}\}$ are independent $\mathcal{N}\left(0, \sigma_{Tst}^{2}\right)$ random variables, with known design variances σ_{Tst}^{2};
- the sampling errors $\{e_{st}^{M}\}$ are independent $\mathcal{N}\left(0, \sigma_{Mst}^{2}\right)$ random variables, with known design variances σ_{Mst}^{2};
- the irregular terms $\{\nu_{st}\}$, representing real variation not explained by the regular trend plus seasonal pattern, are independent and identically distributed (iid) $\mathcal{N}(0, \psi)$ random variables, with unknown variance ψ;
- $\{e_{st}^{T}\}$, $\{e_{st}^{M}\}$ and $\{\nu_{st}\}$ are mutually independent.

The assumed independence of the sampling errors is justified by independent samples drawn state-to-state and wave-to-wave, and the assumed normality is justified by central limiting effects of moderate to large-size stratified samples in each state and wave. Further, we assume that because the mail and telephone surveys are selected and conducted independently, the sampling errors $\{e_{st}^T\}$ and $\{e_{st}^M\}$ are independent of one another. We use simulation to assess the sensitivity of some of our methods to the normality assumption on the random effects in Section 4.4.1 below. The design variances $\{\sigma_{Tst}^2\}$ and $\{\sigma_{Mst}^2\}$ are on the log scale, while the available design variance estimates $\{\widehat{V}_{Tst}\}$ and $\{\widehat{V}_{Mst}\}$ are on the original scale; we address this discrepancy in Section 4.2.2 below.

4.2.2 Design Variance Model

Under the log-normal effort models (4.2.1), the variances of the sampling errors are given by

$$V_{Tst} = \text{Var}\left(\exp(\widehat{T}_{st}) \mid T_{st}\right) = \left\{\exp(\sigma_{Tst}^2) - 1\right\} \exp\left\{2T_{st} + \sigma_{Tst}^2\right\} \tag{4.2.2}$$

and

$$V_{Mst} = \text{Var}\left(\exp(\widehat{M}_{st}) \mid M_{st}\right) = \left\{\exp(\sigma_{Mst}^2) - 1\right\} \exp\left\{2M_{st} + \sigma_{Mst}^2\right\}. \tag{4.2.3}$$

We need to estimate σ_{Tst}^2 and σ_{Mst}^2, incorporating the approximately design-unbiased estimates $\widehat{V}_{Tst}$ and $\widehat{V}_{Mst}$ of V_{Tst} and V_{Mst}, respectively.

We follow an approach related closely to generalized variance function estimation (e.g., Ch. 7 of Wolter (2007)). Assume that given T_{st} and M_{st}, the empirical coefficients of variation (CV's) are log-normally distributed, independent of the effort estimates $\widehat{T}_{st}$ and $\widehat{M}_{st}$:

$$\ln\left(\frac{\widehat{V}_{Tst}}{\exp(2\widehat{T}_{st})}\right) = \boldsymbol{d}_{Tst}^{\intercal}\boldsymbol{\delta}_0^T + \delta_1^T \ln(n_{Tst}) + \eta_{st}^T, \quad \eta_{st}^T \sim \mathcal{N}(0, \tau_T^2) \tag{4.2.4}$$

where $\boldsymbol{d}_{Tst}$ is a vector of known covariates (including state, wave, and state by wave interaction), and

$$\ln\left(\frac{\widehat{V}_{Mst}}{\exp(2\widehat{M}_{st})}\right) = \boldsymbol{d}_{Mst}^{\intercal}\boldsymbol{\delta}_0^M + \delta_1^M \ln(n_{Mst}) + \eta_{st}^M, \quad \eta_{st}^M \sim \mathcal{N}(0, \tau_M^2), \tag{4.2.5}$$

where $\boldsymbol{d}_{Mst}$ is a vector of known covariates. These models can be rewritten as regression models for the design variance estimates, with known offsets:

$$\ln\left(\widehat{V}_{Tst}\right) = 2\widehat{T}_{st} + \boldsymbol{d}_{Tst}^{\intercal}\boldsymbol{\delta}_0^{T} + \delta_1^{T}\ln(n_{Tst}) + \eta_{st}^{T}, \quad \eta_{st}^{T} \sim \mathcal{N}(0, \tau_T^2)$$

and

$$\ln\left(\widehat{V}_{Mst}\right) = 2\widehat{M}_{st} + \boldsymbol{d}_{Mst}^{\intercal}\boldsymbol{\delta}_0^{M} + \delta_1^{M}\ln(n_{Mst}) + \eta_{st}^{M}, \quad \eta_{st}^{M} \sim \mathcal{N}(0, \tau_M^2).$$

Empirically, each of these models fits very well: 94.54% adjusted R^2 value for telephone, and 98.01% adjusted R^2 value for mail.

These empirical models may be of independent interest as generalized variance functions for variance estimation on the original scale: by plugging the point estimate, state, wave, and sample size into the fitted versions of (4.2.4) or (4.2.5), one obtains excellent point estimates of the coefficient of variation.

Assuming that $\widehat{V}_{Tst}$ is exactly unbiased for V_{Tst}, we then have from the log-normal CV model (4.2.4) and the assumed conditional independence of $\widehat{V}_{Tst}$ and $\widehat{T}_{st}$ given T_{st} that

$$\begin{aligned}
&\exp\left\{\boldsymbol{d}_{Tst}^{\intercal}\boldsymbol{\delta}_0^{T} + \delta_1^{T}\ln(n_{Tst}) + \frac{\tau_T^2}{2}\right\} = \mathrm{E}\left[\left.\frac{\widehat{V}_{Tst}}{\exp\left(2\widehat{T}_{st}\right)}\right| T_{st}\right] \\
&= \mathrm{E}\left[\widehat{V}_{Tst} \mid T_{st}\right]\mathrm{E}\left[\exp\left(-2\widehat{T}_{st}\right) \mid T_{st}\right] \\
&= V_{Tst}\exp\left(-2T_{st} + 2\sigma_{Tst}^2\right), \qquad (4.2.6)
\end{aligned}$$

and similarly

$$\begin{aligned}
&\exp\left\{\boldsymbol{d}_{Mst}^{\intercal}\boldsymbol{\delta}_0^{M} + \delta_1^{M}\ln(n_{Mst}) + \frac{\tau_M^2}{2}\right\} = \mathrm{E}\left[\left.\frac{\widehat{V}_{Mst}}{\exp\left(2\widehat{M}_{st}\right)}\right| M_{st}\right] \\
&= \mathrm{E}\left[\widehat{V}_{Mst} \mid M_{st}\right]\mathrm{E}\left[\exp\left(-2\widehat{M}_{st}\right) \mid M_{st}\right] \\
&= V_{Mst}\exp\left(-2M_{st} + 2\sigma_{Mst}^2\right). \qquad (4.2.7)
\end{aligned}$$

Thus, we have from (4.2.2) and (4.2.6) that

$$
\begin{aligned}
&\exp\left\{\boldsymbol{d}_{Tst}^{\intercal}\boldsymbol{\delta}_0^T + \delta_1^T \ln(n_{Tst}) + \frac{\tau_T^2}{2}\right\} \\
&= \left\{\exp(\sigma_{Tst}^2) - 1\right\} \exp\left\{2T_{st} + \sigma_{Tst}^2\right\} \exp\left(-2T_{st} + 2\sigma_{Tst}^2\right) \\
&= \exp(4\sigma_{Tst}^2) - \exp\left(3\sigma_{Tst}^2\right) \qquad (4.2.8)
\end{aligned}
$$

and from (4.2.3) and (4.2.7) that

$$
\begin{aligned}
&\exp\left\{\boldsymbol{d}_{Mst}^{\intercal}\boldsymbol{\delta}_0^M + \delta_1^M \ln(n_{Mst}) + \frac{\tau_M^2}{2}\right\} \\
&= \left\{\exp(\sigma_{Mst}^2) - 1\right\} \exp\left\{2M_{st} + \sigma_{Mst}^2\right\} \exp\left(-2M_{st} + 2\sigma_{Mst}^2\right) \\
&= \exp(4\sigma_{Mst}^2) - \exp\left(3\sigma_{Mst}^2\right). \qquad (4.2.9)
\end{aligned}
$$

The left-hand-side parameters of (4.2.8) can be estimated from (4.2.4) and the left-hand-side parameters of (4.2.9) can be estimated from (4.2.5). The resulting estimates of σ_{Tst}^2 and σ_{Mst}^2 can then be obtained by solving the equations (4.2.8) and (4.2.9), which are quartic polynomials in $\exp(\sigma_{Tst}^2)$ and $\exp(\sigma_{Mst}^2)$. Using Descartes' rule of signs, it can be shown that each of these quartic equations has one negative real root, two complex conjugate roots, and one positive real root. The solutions for σ_{Tst}^2 and σ_{Mst}^2 are then the logarithms of the unique, positive real roots, which can be obtained via standard numerical procedures. While these solutions are in fact estimates, we will treat them as fixed and known in what follows, as is standard in the small area estimation techniques which we will apply in subsequent sections.

The resulting design variances on the log scale, σ_{Tst}^2 and σ_{Mst}^2, are strongly correlated with the variance approximations from Taylor linearization, $\widehat{V}_{Tst}\exp\left(-2\widehat{T}_{st}\right)$ and $\widehat{V}_{Mst}\exp\left(-2\widehat{M}_{st}\right)$: 0.798 and 0.803, respectively. But they are not identical (see Figure 4.1), and the method described forces analytical consistency between the mean model and the variance model.

4.2.3 Fay-Herriot Small Area Estimation Model

Define

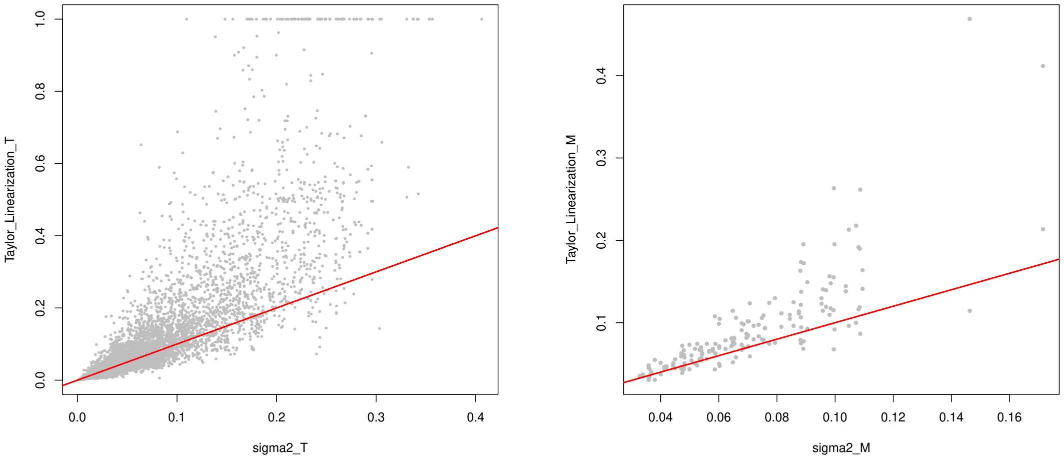

Figure 4.1: Estimated design variances for log-effort via Taylor linearization versus solution of the quartic polynomial equations (4.2.8) for telephone (left panel) and (4.2.9) for mail (right panel).

$$\boldsymbol{x}_{st}^{\top} = \begin{cases} \boldsymbol{x}_{Tst}^{\top}, & \text{if no mail estimate is available;} \\ \boldsymbol{x}_{Mst}^{\top}, & \text{if no telephone estimate is available;} \\ (\boldsymbol{x}_{Tst} + \boldsymbol{x}_{Mst})^{\top}/2, & \text{otherwise.} \end{cases}$$

Then it is convenient to write

$$\begin{aligned} Y_{st} &= \begin{cases} \widehat{T}_{st}, & \text{if no mail estimate is available;} \\ \widehat{M}_{st}, & \text{if no telephone estimate is available;} \\ \left(\widehat{T}_{st} + \widehat{M}_{st}\right)/2, & \text{otherwise;} \end{cases} \\ &= \begin{cases} \boldsymbol{x}_{Tst}^{\top}\boldsymbol{\beta} + \nu_{st} + e_{st}^{T}, & \text{if no mail estimate is available;} \\ \boldsymbol{x}_{Mst}^{\top}\boldsymbol{\beta} + \nu_{st} + e_{st}^{M}, & \text{if no telephone estimate is available;} \\ (\boldsymbol{x}_{Tst} + \boldsymbol{x}_{Mst})^{\top}\boldsymbol{\beta}/2 + \nu_{st} + (e_{st}^{T} + e_{st}^{M})/2, & \text{otherwise;} \end{cases} \\ &= \boldsymbol{x}_{st}^{\top}\boldsymbol{\beta} + \nu_{st} + e_{st}. \end{aligned} \tag{4.2.10}$$

This model then follows exactly the linear mixed model structure of Fay & Herriot (1979), with direct estimates Y_{st} equal to regression model plus random effect ν_{st} plus sampling error with "known" design variance, given by

$$D_{st} = \begin{cases} \sigma^2_{Tst}, & \text{if no mail estimate is available;} \\ \sigma^2_{Mst}, & \text{if no telephone estimate is available;} \\ \frac{1}{4}\left(\sigma^2_{Tst} + \sigma^2_{Mst}\right), & \text{otherwise.} \end{cases}$$

Averaging the telephone and mail estimates results in a small loss of information, since we are replacing two correlated observations with one observation, but allows the use of standard software for estimation.

4.3 Methods

4.3.1 Estimation for the Fay-Herriot Model

Define $\mathcal{A} = \{(s,t) : Y_{st} \text{ is not missing}\}$ to be the set of all state by year-wave combinations for which we have an estimate from either survey. Let m denote the size of the set $\mathcal{A}$. Define $\boldsymbol{X} := [\boldsymbol{x}^{\intercal}_{st}]_{(s,t)\in\mathcal{A}}$, $\boldsymbol{Y} := [Y_{st}]_{(s,t)\in\mathcal{A}}$. We have

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + [\nu_{st}]_{(s,t)\in\mathcal{A}} + [e_{st}]_{(s,t)\in\mathcal{A}}.$$

Then $\boldsymbol{\Sigma}(\psi) := \text{Var}\,(\boldsymbol{Y}) = \text{diag}\{\psi + D_{st}\}_{(s,t)\in\mathcal{A}}$. If ψ were known, the best linear unbiased estimator (BLUE) of $\boldsymbol{\beta}$ would be

$$\widetilde{\boldsymbol{\beta}}_{\psi} = \left\{\boldsymbol{X}^{\intercal}\boldsymbol{\Sigma}^{-1}(\psi)\boldsymbol{X}\right\}^{-1}\boldsymbol{X}^{\intercal}\boldsymbol{\Sigma}^{-1}(\psi)\boldsymbol{Y}. \tag{4.3.1}$$

Since ψ is not known, we replace it by a consistent estimator to obtain

$$\widehat{\boldsymbol{\beta}} = \left\{\boldsymbol{X}^{\intercal}\boldsymbol{\Sigma}^{-1}(\hat{\psi})\boldsymbol{X}\right\}^{-1}\boldsymbol{X}^{\intercal}\boldsymbol{\Sigma}^{-1}(\hat{\psi})\boldsymbol{Y}. \tag{4.3.2}$$

We will use the Restricted Maximum Likelihood (REML) estimate $\hat{\psi}$ unless otherwise indicated.

4.3.2 Prediction

In the classical Fay-Herriot context, it is of interest to predict

$$\boldsymbol{x}_{st}^{\mathsf{T}}\boldsymbol{\beta} + \nu_{st}$$

from (4.2.10). In our setting, however, we seek to predict

$$\phi_{st} = \boldsymbol{z}_{st}^{\mathsf{T}}\boldsymbol{\beta} + \nu_{st}, \tag{4.3.3}$$

where $\boldsymbol{z}_{st}$ may not equal $\boldsymbol{x}_{st}$. For example, for a past time point with a telephone survey estimate but no mail survey estimate, we may want to use

$$\boldsymbol{z}_{st}^{\mathsf{T}} = \boldsymbol{x}_{Mst}^{\mathsf{T}} = [\boldsymbol{a}_{st}^{\mathsf{T}}, \boldsymbol{b}_{st}^{\mathsf{T}}, \mathbf{0}^{\mathsf{T}}]$$

to predict the mail target M_{st}, while for a future time point with a mail survey estimate but no telephone, we may want to use

$$\boldsymbol{z}_{st}^{\mathsf{T}} = [\boldsymbol{a}_{st}^{\mathsf{T}}, \mathbf{0}^{\mathsf{T}}, \mathbf{0}^{\mathsf{T}}]$$

to predict the telephone target, corrected for the wireless effect: $T_{st} - w_{st}\boldsymbol{c}_{st}^{\mathsf{T}}\boldsymbol{\gamma} = \boldsymbol{a}_{st}^{\mathsf{T}}\boldsymbol{\alpha} + \nu_{st}$.

Let $\boldsymbol{\lambda}_{st}$ denote an $m \times 1$ vector with a one in the (s,t)th position and zero elsewhere. Under normality, it is well-known that the best mean square predictor of ϕ_{st} in (4.3.3) is

$$\phi_{st}\left(\boldsymbol{\beta}, \psi\right) = \boldsymbol{z}_{st}^{\mathsf{T}}\boldsymbol{\beta} + \psi\boldsymbol{\lambda}_{st}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\psi)(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}), \tag{4.3.4}$$

which is feasible only if both $\boldsymbol{\beta}$ and ψ are both known. If only ψ is known, the best linear unbiased predictor (BLUP)

$$\phi_{st}\left(\widetilde{\boldsymbol{\beta}}_{\psi},\psi\right)=\boldsymbol{z}_{st}^{\intercal}\widetilde{\boldsymbol{\beta}}(\psi)+\psi\boldsymbol{\lambda}_{st}^{\intercal}\boldsymbol{\Sigma}^{-1}(\psi)(\boldsymbol{Y}-\boldsymbol{X}\widetilde{\boldsymbol{\beta}}(\psi)) \tag{4.3.5}$$

is obtained by plugging the BLUE from (4.3.1) into (4.3.4). Finally, if neither $\boldsymbol{\beta}$ nor ψ is known, then the empirical best linear unbiased predictor (EBLUP) can be obtained by substituting a consistent estimator of ψ into (4.3.5):

$$\phi_{st}\left(\widehat{\boldsymbol{\beta}},\hat{\psi}\right)=\boldsymbol{z}_{st}^{\intercal}\widehat{\boldsymbol{\beta}}+\hat{\psi}\boldsymbol{\lambda}_{st}^{\intercal}\boldsymbol{\Sigma}^{-1}(\hat{\psi})(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}), \tag{4.3.6}$$

where $\widehat{\boldsymbol{\beta}}$ is given by (4.3.2). These EBLUPs are the proposed calibrated values on the log scale.

4.3.3 Mean Square Error Approximation

To assess the uncertainty of the calibrated values, we adapt the approach of Datta & Lahiri (2000) in approximating the MSE of $\phi_{st}\left(\widehat{\boldsymbol{\beta}},\hat{\psi}\right)$. It can be shown that

$$\begin{aligned}\text{MSE}\left\{\phi_{st}\left(\widehat{\boldsymbol{\beta}},\hat{\psi}\right)\right\} &= \text{E}\left[\left\{\phi_{st}\left(\widehat{\boldsymbol{\beta}},\hat{\psi}\right)-\phi_{st}\right\}^{2}\right]\\ &= \text{E}\left[\left\{\phi_{st}\left(\boldsymbol{\beta},\psi\right)-\phi_{st}\right\}^{2}\right]+\text{E}\left[\left\{\phi_{st}\left(\widetilde{\boldsymbol{\beta}}_{\psi},\psi\right)-\phi_{st}\left(\boldsymbol{\beta},\psi\right)\right\}^{2}\right]\\ &\quad+\text{E}\left[\left\{\phi_{st}\left(\widehat{\boldsymbol{\beta}},\hat{\psi}\right)-\phi_{st}\left(\widetilde{\boldsymbol{\beta}}_{\psi},\psi\right)\right\}^{2}\right]\\ &= \dot{g}_{1st}(\psi)+\dot{g}_{2st}(\psi)+\dot{g}_{3st}(\psi)+o\left(m^{-1}\right),\end{aligned} \tag{4.3.7}$$

where

$$\dot{g}_{1st}(\psi)=\frac{\psi D_{st}}{\psi+D_{st}},$$

$$\dot{g}_{2st}(\psi)=\left(\frac{\psi(\boldsymbol{z}_{st}-\boldsymbol{x}_{st})^{\intercal}+D_{st}\boldsymbol{z}_{st}^{\intercal}}{\psi+D_{st}}\right)\left[\sum_{u\in\mathcal{A}}(\psi+D_{u})^{-1}\boldsymbol{x}_{u}\boldsymbol{x}_{u}^{\intercal}\right]^{-1}\left(\frac{\psi(\boldsymbol{z}_{st}-\boldsymbol{x}_{st})^{\intercal}+D_{st}\boldsymbol{z}_{st}^{\intercal}}{\psi+D_{st}}\right)^{\intercal},$$

and

$$\dot{g}_{3st}(\psi) = \frac{2D_{st}^2}{(\psi + D_{st})^3} \frac{1}{\sum_{u \in \mathcal{A}} (\psi + D_u)^{-2}}.$$

The terms $\dot{g}_{1st}(\psi)$ and $\dot{g}_{3st}(\psi)$ are identical to the terms $g_{1st}(\psi)$ and $g_{3st}(\psi)$ in section 4 of Datta & Lahiri (2000), while $\dot{g}_{2st}(\psi)$ simplifies to $g_{2st}(\psi)$ of that paper in the special case of $\boldsymbol{z}_{st} = \boldsymbol{x}_{st}$. We defer the proofs to the Appendix.

4.3.4 Mean Square Error Estimation

We now propose an estimator of the MSE approximation in (4.3.7). Using arguments like those in section 5 of Datta & Lahiri (2000), it can be shown that

$$\begin{aligned}
\mathrm{E}\left[\dot{g}_{1st}(\hat{\psi})\right] &\simeq \dot{g}_{1st}(\psi) - \dot{g}_{3st}(\psi) \\
\mathrm{E}\left[\dot{g}_{2st}(\hat{\psi})\right] &\simeq \dot{g}_{2st}(\psi) \\
\mathrm{E}\left[\dot{g}_{3st}(\hat{\psi})\right] &\simeq \dot{g}_{3st}(\psi)
\end{aligned}$$

and hence an approximately unbiased estimator of the MSE approximation in (4.3.7) is given by

$$\mathrm{mse}\left\{\phi_{st}\left(\widehat{\boldsymbol{\beta}}, \hat{\psi}\right)\right\} = \dot{g}_{1st}(\hat{\psi}) + \dot{g}_{2st}(\hat{\psi}) + 2\dot{g}_{3st}(\hat{\psi}). \tag{4.3.8}$$

We assess the quality of the asymptotic approximation (4.3.7) and its estimator (4.3.8) via simulation in Section 4.4.1.

4.3.5 Prediction on the Original Scale

To compute predictors on the original scale, we back-transform by exponentiating the EBLUP from (4.3.6) and adjust for the nonlinearity of the back-transformation using the estimated MSE from (4.3.8):

$$\widehat{\exp(\phi_{st})} = \exp\left[\phi_{st}\left(\widehat{\boldsymbol{\beta}}, \hat{\psi}\right) + \frac{1}{2}\mathrm{mse}\left\{\phi_{st}\left(\widehat{\boldsymbol{\beta}}, \hat{\psi}\right)\right\}\right], \tag{4.3.9}$$

which is an estimator of the best mean square predictor under the normal model, and a standard adjustment even without the normality assumption.

4.3.6 Moving Average of the Predictions

One way to get smoother predictions is to compute a moving average of the EBLUPs in (4.3.6). Here we denote the new predictor as $\widehat{\phi}_{st\mathrm{MA}}$. As we mention in Section 4.1, t denotes year-wave. In this subsection we separate t as t_1 and t_2 for year and wave respectively. Then

$$\widehat{\phi}_{st\mathrm{MA}} = \sum_{j=-K}^{K} a_j \phi_{s(t_1+j)t_2}\left(\widehat{\boldsymbol{\beta}}, \hat{\psi}\right), \tag{4.3.10}$$

where $\{a_j\}$ is a sequence of constants with $\sum_{j=-K}^{K} a_j = 1$. We apply this filter to every wave of each state across different years. To deal with edge effects at year 2017, we average five waves for years up to 2013 as $\sum_{j=0}^{4} \phi_{s(t_1+j)t_2}\left(\widehat{\boldsymbol{\beta}}, \hat{\psi}\right)/5$; we average four waves for year 2014 as $\sum_{j=0}^{3} \phi_{s(t_1+j)t_2}\left(\widehat{\boldsymbol{\beta}}, \hat{\psi}\right)/4$; three waves for year 2015; two waves for 2016; and a single wave for 2017 (that is, 2017 is left unchanged). The MSE is given as

$$\begin{aligned}\mathrm{MSE}\left(\widehat{\phi}_{st\mathrm{MA}}\right) = &\sum_{j=-K}^{K} a_j^2 \left\{\mathrm{MSE}\left(\widehat{\phi}_{s(t_1+j)t_2}\right) + \mathrm{E}\left[\left(\phi_{s(t_1+j)t_2} - \phi_{st_1t_2}\right)^2\right]\right\} \\ &+ \sum\sum_{i\neq j} a_i a_j \mathrm{E}\left[\left(\phi_{s(t_1+i)t_2} - \phi_{st_1t_2}\right)\left(\phi_{s(t_1+j)t_2} - \phi_{st_1t_2}\right)\right] + \mathcal{O}\left(m^{-1/2}\right).\end{aligned} \tag{4.3.11}$$

By plugging in the EBLUPs, we have

$$\begin{aligned}\mathrm{mse}\left\{\widehat{\phi}_{st\mathrm{MA}}\right\} = &\sum_{j=-K}^{K} a_j^2 \left\{\mathrm{mse}\left(\widehat{\phi}_{s(t_1+j)t_2}\right) + \left(\widehat{\phi}_{s(t_1+j)t_2} - \widehat{\phi}_{st_1t_2}\right)^2\right\} \\ &+ \sum\sum_{i\neq j} a_i a_j \left(\widehat{\phi}_{s(t_1+i)t_2} - \widehat{\phi}_{st_1t_2}\right)\left(\widehat{\phi}_{s(t_1+j)t_2} - \widehat{\phi}_{st_1t_2}\right).\end{aligned} \tag{4.3.12}$$

Once we have the new predictors, we transform them back to the original scale as shown in (4.3.9). Plots of the moving averaged predictions are omitted. Derivation of (4.3.11) is deferred to the Appendix.

4.4 Empirical Results

4.4.1 Simulation

In this section, we investigate the performance of our second-order approximation of MSE and the estimated MSE under a setting that mimics the reconciliation problem of this paper, but with a smaller number of observed time points: 17 states and six years (1985, 1995, 2005, 2010, 2015, and 2016) of six waves each, with telephone effort estimates for all waves, and with mail effort estimates for only the final two years. In this setting, $m =$ (17 states)(6 waves)(6 years) $= 612$. We take the wireless values and US Census population counts from the actual data.

We use the estimates from model (4.2.10) fitted to shore data, with intercept, log(population), state indicators, wave indicators, state by log(population) interaction, and state by wave; plus wireless and its interactions with log(population), state indicators, and wave indicators; plus an indicator for presence of a mail survey estimate and the mail indicator's interactions with log(population), state indicators, and wave indicators. We pick $\psi = 0.11$, again from the fit of the model. The simulation model is similar to the final model selected in Section 4.4.2 below.

We consider three different patterns for the design variances $\{D_{st}\}$. First, we sample six actual design variances for each simulated state, arrange the six into a "peaked" seasonal pattern, and replicate this seasonal pattern across all six years to create pattern (b). We consider two additional settings, by multiplying pattern (b) by 0.5 to yield pattern (a), and multiplying pattern (b) by 2.0 to yield pattern (c). The simulated sampling errors $\{e_{st}\}$ in (4.2.10) are then generated independently as $\mathcal{N}(0, D_{st})$ under each pattern.

Following Datta et al. (2005), we consider three distributions to simulate the normalized random effects:

- $\{\psi^{-1/2}\nu_{st}\}$ iid $\mathcal{N}(0, 1)$;

- $\{\psi^{-1/2}\nu_{st}\}$ iid Laplace$(0, 1/\sqrt{2})$;
- $\{\psi^{-1/2}\nu_{st}\}$ iid centered Exponential(1) (that is, exponential random variables centered to mean zero).

Under each distribution, $\mathrm{E}\left[\nu_{st}\right] = 0$ and $\mathrm{Var}\left(\nu_{st}\right) = \psi$.

For each combination of sampling variance pattern and random effect distribution, we generate 1000 data sets from model (4.2.10). For each simulated data set, we use the `R` package `sae` Molina & Marhuenda (2015) to compute $\hat{\psi}$ via REML and $\widehat{\boldsymbol{\beta}}$. We compute the EBLUPs in (4.3.6) for the mail targets $\{M_{st}\}$, approximate their MSEs using (4.3.7), and estimate their MSEs using (4.3.8). We then compare the approximations and the estimates to the true (Monte Carlo) MSEs over the 1000 simulated realizations.

Figure 4.2 shows plots of the MSE approximation and the estimated MSE versus the true MSE for each of the nine simulation scenarios. Here the gray dots are the MSE approximations and the black circles are the estimated MSE's. The approximations and estimates are nearly overlapping in all cases, indicating that the MSE estimates are essentially unbiased for the MSE approximations. Further, the points are all very close to the (0,1) reference line, indicating that the proposed methodology yields acceptable MSE estimates across a range of settings.

4.4.2 Calibration of the CHTS and FES Estimates

For the data described in Section 4.1, we use the `R` package `sae` (Molina & Marhuenda, 2015) to fit a number of models via maximum likelihood for both shore fishing and private boat fishing, and compare the models via their AIC values. The smallest model considered includes intercept, log(population), state indicators, wave indicators, state by log(population) interaction, and state by wave interaction. That is, the smallest model includes no differences due to survey methodology and instead drops the terms $\boldsymbol{b}_{st}^{\mathsf{T}}\boldsymbol{\mu}$ and $w_{st}\boldsymbol{c}_{st}^{\mathsf{T}}\gamma$ from (4.2.1). The largest model considered adds wireless and its interactions with log(population), state indicators, wave indicators, and state by log(population), together with an indicator for presence of a mail survey estimate and the mail indicator's interactions with log(population), state indicators, and wave indicators. The omission

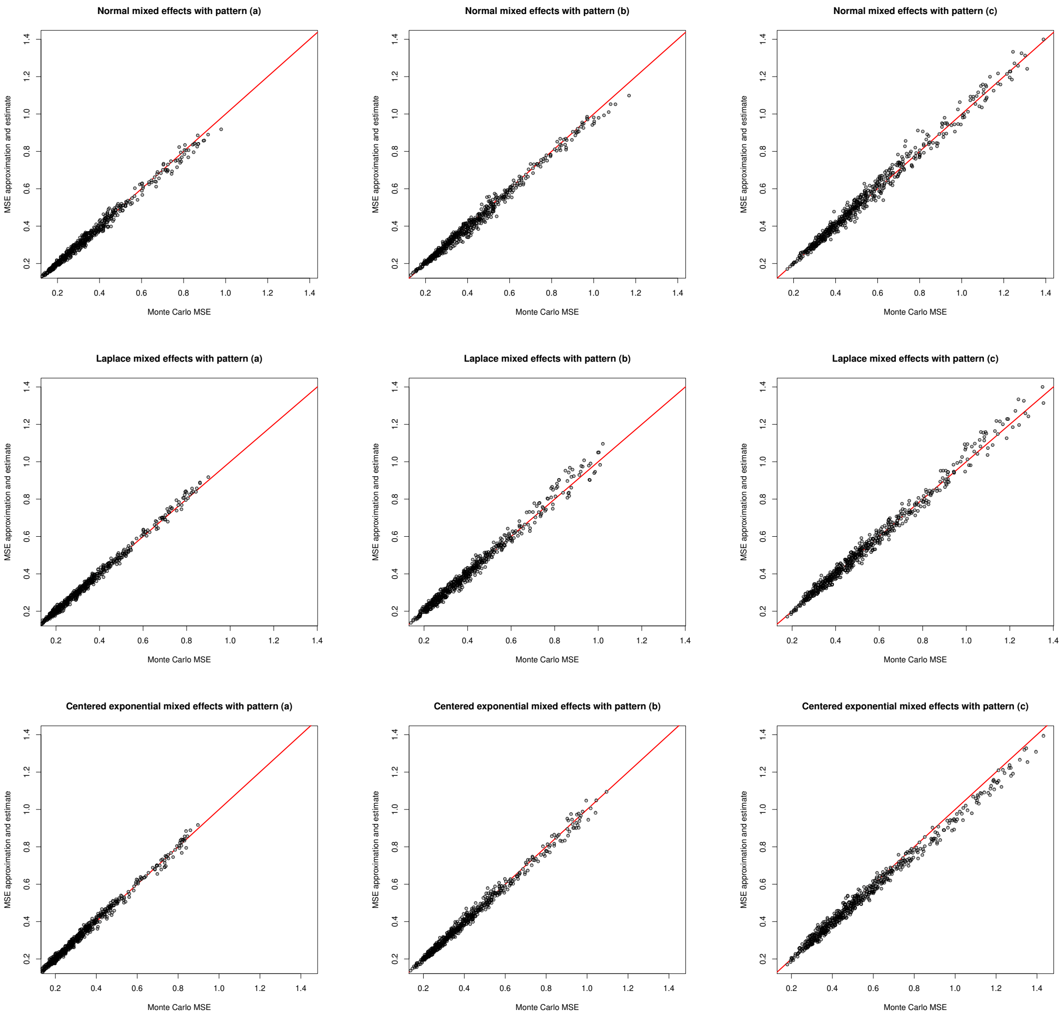

Figure 4.2: MSE approximation (solid gray dots) and estimated MSE's (open black circles) versus true MSE from Monte Carlo, for random effect distributions normal, Laplace, and centered exponential across the rows, and sampling error patterns (a), (b), and (c) across the columns.

of the higher order interactions between wireless and the mail indicator is due to parsimony: for the mail indicator in particular, there are only 17 states and 11 waves from which to estimate the parameters $\boldsymbol{\mu}$ in model (4.2.1).

Numerous submodels between the smallest and largest are considered; the best five models and additional reference models are given in Table 4.1 for shore fishing and Table 4.2 for private boat fishing. The tables are ordered by AIC values, with the best models at the top. The models that ignore some (largest minus all mail, largest minus all wireless) or all (smallest) of the survey mode differences are not competitive with the models that include these factors. The largest model considered is quite competitive, with the best models dropping a small number of interactions from that largest model.

While not the best model for either shore or private boat, the largest model minus the mail by log(population) and mail by state interaction is fifth best in both cases. It is operationally convenient to use a common model for both reconciliations, and this particular model is further convenient because, when extrapolating back in time, it involves only wave level shifts once the effect of wireless has died out. We also use the first two waves of 2018 as our out of sample data for prediction. With the AIC versus MSE of prediction plot (not shown here), this model is obviously among the top a few. We therefore choose this model as the final model for both modes of fishing, and refit it using REML to estimate the unknown variance ψ. We then compute EBLUPs of the mail target $\{M_{st}\}$ for all states and waves.

An example for Alabama shore fishing is shown in Figure 4.3 and an example for Florida private boat fishing is shown in Figure 4.4. In each figure, we show the raw effort survey estimates and the EBLUPs. The EBLUPs can be seen as a smoothed version of the estimates adjusted for mail methodology and wireless effects.

4.5 Discussion

The proposed methodology accounts for various sources of variation in the effort series from each survey, including trend, seasonality and irregular terms in the true effort series, together with

Table 4.1: Maximized log(likelihood), AIC and residual degrees of freedom (df) for various models fitted to effort estimates for shore fishing. See text for description of largest model.

Model is largest minus terms below:	log(likelihood)	AIC	df
mail:log(pop), mail:state, wireless:wave	-2129.14	4564.28	3022
mail:state, wireless:wave	-2128.35	4564.69	3021
mail:log(pop) and wireless:wave	-2113.43	4564.86	3006
wireless:wave	-2113.42	4566.85	3005
mail:log(pop) and mail:state	-2127.22	4570.45	3017
nothing (largest)	-2111.64	4573.28	3000
mail interactions	-2137.25	4580.51	3022
wireless interactions	-2223.52	4719.05	3038
all interactions	-2256.42	4742.84	3050
all wireless	-2243.37	4758.73	3029
all mail	-2267.37	4838.73	3023
all mail and all wireless (smallest)	-2440.35	5106.70	3052

Table 4.2: Maximized log(likelihood), AIC and residual degrees of freedom (df) for various models fitted to effort estimates for private boat fishing. See text for description of largest model.

Model is largest minus terms below:	log(likelihood)	AIC	df
nothing (largest)	-1482.28	3314.55	2990
mail:log(pop)	-1483.28	3314.56	2991
mail:log(pop) and wireless:wave	-1489.21	3316.42	2996
wireless:wave	-1488.23	3316.47	2995
mail:log(pop) and mail:state	-1503.36	3322.73	3007
mail:state	-1502.50	3323.00	3006
mail interactions	-1528.13	3362.27	3012
all mail	-1598.62	3501.23	3013
wireless interactions	-1623.17	3520.33	3028
all interactions	-1708.39	3646.78	3050
all wireless	-1739.02	3750.03	3029
all mail and all wireless (smallest)	-1837.91	3901.82	3052

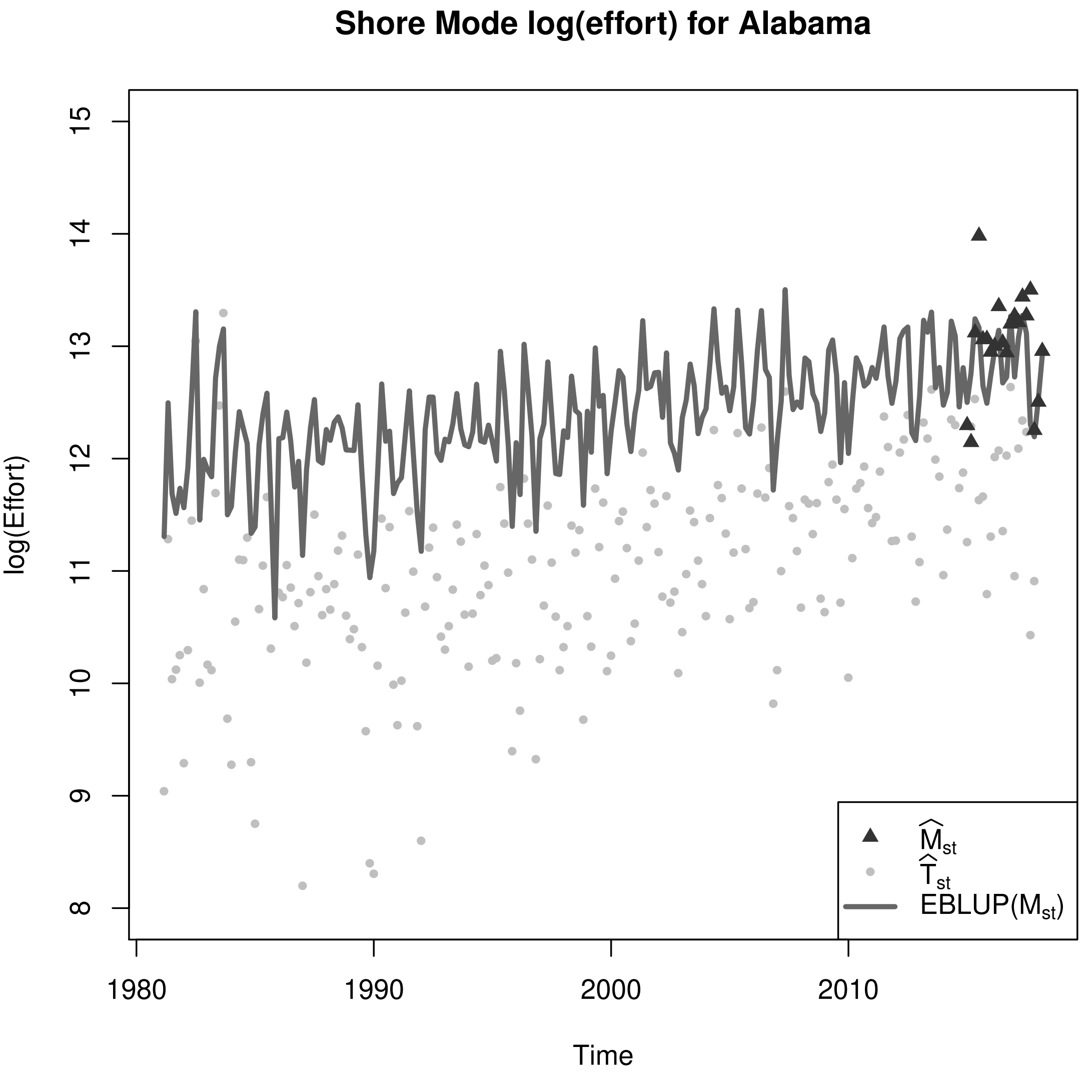

Figure 4.3: EBLUPs $\left\{\phi_{st}\left(\widehat{\boldsymbol{\beta}}, \hat{\psi}\right)\right\}$ (curve) of mail targets $\{M_{st}\}$ for shore fishing log-effort in Alabama. Gray dots are telephone log-effort estimates $\{\widehat{T}_{st}\}$ and black triangles are mail log-effort estimates $\{\widehat{M}_{st}\}$.

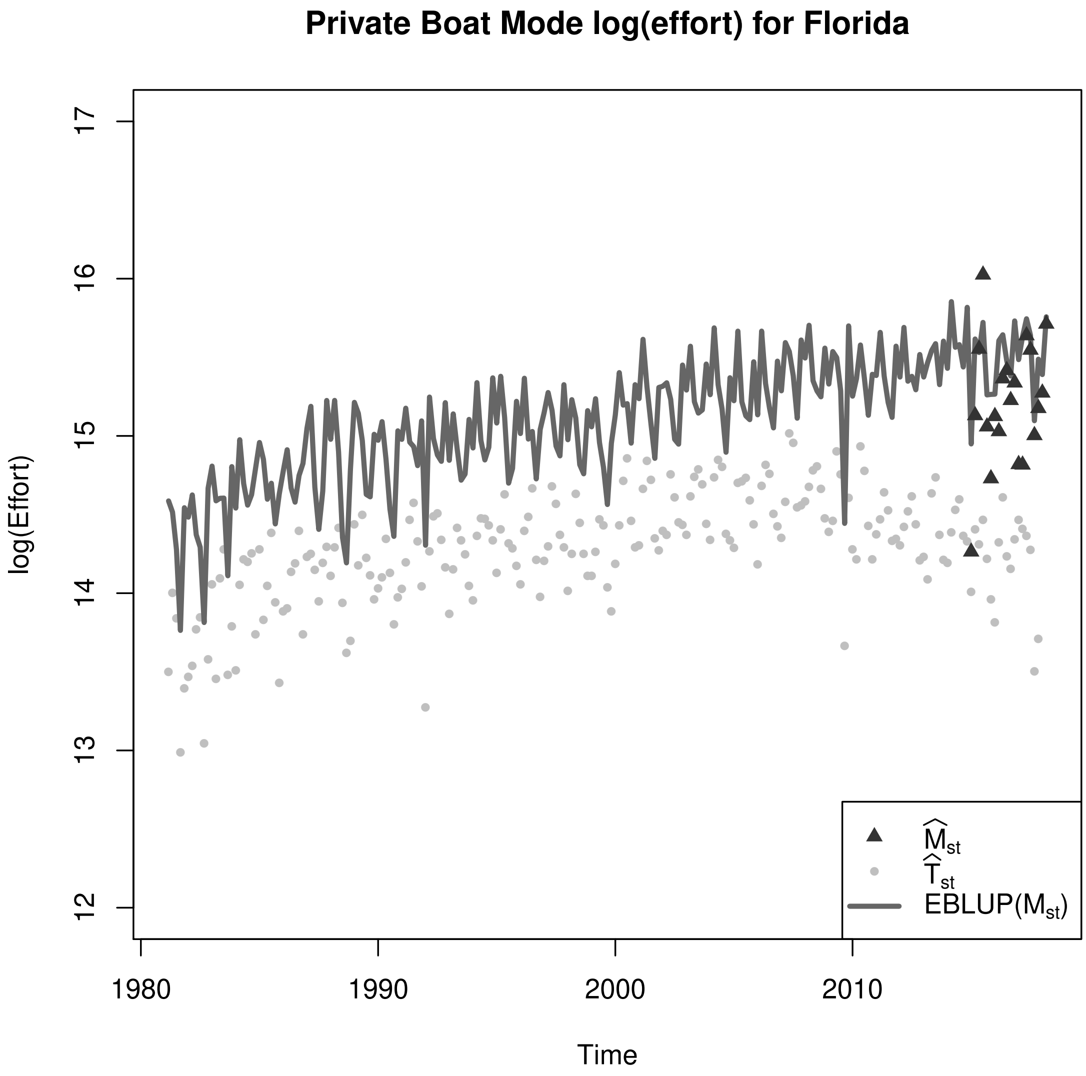

Figure 4.4: EBLUPs $\left\{\phi_{st}\left(\widehat{\boldsymbol{\beta}}, \hat{\psi}\right)\right\}$ (curve) of mail targets $\{M_{st}\}$ for private boat fishing in Florida. Gray dots are telephone log-effort estimates $\{\widehat{T}_{st}\}$ and black triangles are mail log-effort estimates $\{\widehat{M}_{st}\}$.

survey mode effects in the two series. The model assumes that differences in measurement and nonresponse errors between the two surveys would be stable over time, while the changes in coverage error over time due to growth in wireless-only households is explicitly modeled. Further, the methodology accounts for uncertainty due to sampling error, using a novel approach to ensure analytical consistency in mapping design variances estimated on the original scale to design variances estimated on the log scale.

As formulated in this paper, the reconciliation methodology turns out to follow a standard, well-established procedure: Fay-Herriot small area estimation. This means that the calibrated values turn out to be empirical best linear unbiased predictors under a linear mixed model fitted using likelihood-based techniques. The method is flexible enough to provide optimal calibrated values for different problems: predicting mail targets using telephone-only data, or predicting telephone targets using mail-only data, for example.

Uncertainty is quantified via a mean square error approximation that adapts existing methods from the literature. Simulation results show that the mean square error approximation and its estimator are highly accurate for the kinds of sample sizes and sampling errors present in the calibration data. The methodology is readily implemented with standard software.

Chapter 5

Conclusion

This dissertation addresses two important problems in the analysis of complex surveys: nonparametric testing for informative selection, and reconciling estimates from two different surveys, with particular application to addressing discontinuities due to survey redesign.

We propose two classes of nonparametric tests for informative selection. Both tests build on the idea of comparing weighted and unweighted characteristics from the same sample, but use this comparison with nonparametric two-sample tests, instead of applying it to parameter estimates. One class of nonparametric tests is based on the difference between weighted and unweighted empirical cdfs. By scaling the difference with an appropriate constant, we prove that the normalized difference converges to a scaled Brownian bridge, and we construct the first class of tests by applying various functionals to the Brownian bridge. We particularly focus on statistics direcly analogous to the classical Kolmogorov–Smirnov statistic and Crámer–Von Mises statistic. We derive their limiting distributions under the null hypothesis of noninformative selection and show via simulation that the tests have correct size under the null and good power across a range of informative alternatives.

The second class of tests is derived by comparing the maximum mean discrepancy between the weighted and unweighted sample in a reproducing kernel Hilbert space. In this setting, we compare weighted and unweighted estimates of probability measures via their maximum mean discrepancy. We derive the asymptotic distribution of the test statistic under the null hypothesis of noninformative selection. Importantly, the test can be conducted with scalar, vector, or other types of responses, provided these random objects are in a Hilbert space. We show via simulation that the tests have correct size under the null and good power across a range of informative alternatives, including in the multivariate case.

Unlike the existing parametric tests that assume correctly-specified models, our methods require no distributional assumptions and no model specification. The computational costs are trivial and generic: the same computations are done regardless of the design, as the tests depend only on the sampling weights and the response. The tests show competitive power relative to existing parametric tests, and in some cases are the only known tests. Thus, it is always worth running our tests first, followed by estimation without adjusting for the design if the test shows no evidence of informative selection, or adjusting for design if the test indicates informative selection. Pre-test estimators that follow this approach have good performance in our limited simulations.

We also provide a small area estimation procedure to reconcile two surveys. By modeling the design variances and applying them within a Fay–Herriot model, we produce the EBLUPs of the variable of interest for new sets of covariates. A second-order approximation of the mean square error of the EBLUPs is derived. With our model, it is convenient to produce estimates for both surveys at any sets of covariates. We anticipate that the need to reconcile different sets of survey estimates will continue, given the rapidly-changing landscape of complex surveys.

References

Altun, Y., & Smola, A. (2006). Unifying divergence minimization and statistical inference via convex duality. In G. Lugosi & H. U. Simon (Eds.), *Learning theory* (pp. 139–153). Berlin, Heidelberg: Springer Berlin Heidelberg.

Andrews, R., Brick, J. M., & Mathiowetz, N. A. (2014). *Development and testing of recreational fishing effort surveys* (Tech. Rep.). Retrieved from `http://www.st.nmfs.noaa.gov/Assets/recreational/pdf/2012-FES_w_review_and_comments_FINAL.pdf`

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society*, *68*(3), 337–404. Retrieved from `http://www.jstor.org/stable/1990404`

Battese, G. E., Harter, R. M., & Fuller, W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, *83*, 28–36.

Berger, Y. G. (1998). Rate of convergence to normal distribution for the horvitz-thompson estimator. *Journal of Statistical Planning and Inference*, *67*(2), 209 - 226. Retrieved from `http://www.sciencedirect.com/science/article/pii/S0378375897001079` doi: https://doi.org/10.1016/S0378-3758(97)00107-9

Bertail, P., Chautru, E., & Clémençon, S. (2013, October). *Empirical processes in survey sampling.* Retrieved from `https://hal.archives-ouvertes.fr/hal-00989585` (Supplementary materials are available for this article.)

Billingsley, P. (1999). *Convergence of probability measures* (Second ed.). New York: John Wiley & Sons Inc. (A Wiley-Interscience Publication)

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review/Revue Internationale de Statistique*, *51*, 279–292.

Blumberg, S., & Luke, J. (2013). *Wireless substitution: Early release of estimates from the National Health Interview Survey, July–December 2012.* (Tech. Rep.). National Center for Health Statistics. Retrieved from `http://www.cdc.gov/nchs/nhis.htm`

Boistard, H., Lopuhaä, H. P., & Ruiz-Gazen, A. (2012). Approximation of rejective sampling inclusion probabilities and application to high order correlations. *Electron. J. Statist.*, *6*, 1967–1983. Retrieved from `https://doi.org/10.1214/12-EJS736` doi: 10.1214/12-EJS736

Boistard, H., Lopuhaä, H. P., & Ruiz-Gazen, A. (2017a, 08). Functional central limit theorems for single-stage sampling designs. *Ann. Statist.*, *45*(4), 1728–1758. Retrieved from `https://doi.org/10.1214/16-AOS1507` doi: 10.1214/16-AOS1507

Boistard, H., Lopuhaä, H. P., & Ruiz-Gazen, A. (2017b). Supplement to 'functional central limit theorems for single-stage samplings designs'.
doi: 10.1214/16-AOS1507SUPP.

Bonnéry, D., Breidt, F. J., & Coquet, F. (2018, 05). Asymptotics for the maximum sample likelihood estimator under informative selection from a finite population. *Bernoulli*, *24*(2), 929–955. Retrieved from `https://doi.org/10.3150/16-BEJ809` doi: 10.3150/16-BEJ809

Breidt, F. J., & Opsomer, J. D. (2000). Local polynomial regression estimators in survey sampling. *Annals of Statistics*, *28*, 1026–1053.

Breslow, N. E., & Cain, K. C. (1988, 03). Logistic regression for two-stage case-control data. *Biometrika*, *75*(1), 11-20. Retrieved from `https://doi.org/10.1093/biomet/75.1.11` doi: 10.1093/biomet/75.1.11

Casella, G., & Berger, R. (2002). *Statistical inference*. Thomson Learning.

Chambers, R., Steel, D., Wang, S., & Welsh, A. (2012). *Maximum likelihood estimation for sample surveys*. CRC Press. Retrieved from `https://books.google.com/books?id=pBTNBQAAQBAJ`

Conti, P. L. (2014, Nov 01). On the estimation of the distribution function of a finite population under high entropy sampling designs, with applications. *Sankhya B*, *76*(2), 234–259. Retrieved from `https://doi.org/10.1007/s13571-014-0083-x` doi: 10.1007/s13571-014-0083-x

Conti, P. L., Marella, D., & Mecatti, F. (2017, 05). Recovering sampling distributions od statistics of finite populations via resampling: a predictive approach.

Cramér, H. (1928). On the composition of elementary errors. *Scandinavian Actuarial Journal*, *1928*(1), 141-180. doi: 10.1080/03461238.1928.10416872

Cressie, N. A. C. (1993). *Statistics for spatial data* (2nd ed.). New York: John Wiley & Sons.

Curtin, R., Presser, S., & Singer, E. (2005). Changes in telephone survey nonresponse over the past quarter century. *Public Opinion Quarterly*, *69*(1), 87. Retrieved from `+http://dx.doi.org/10.1093/poq/nfi002` doi: 10.1093/poq/nfi002

Darling, D. A. (1957). The kolmogorov-smirnov, cramer-von mises tests. *The Annals of Mathematical Statistics*, *28*(4), 823–838. Retrieved from `http://www.jstor.org/stable/2237048`

Datta, G. S., & Lahiri, P. (2000). A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. *Statistica Sinica*, *10*, 613–627.

Datta, G. S., Rao, J. N. K., & Smith, D. D. (2005). On measuring the variability of small area estimators under a basic area level model. *Biometrika*, *92*(1), 183–196.

Deville, J.-C., & Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, *87*, 376–382.

Donsker, M. (1951). *An invariance principle for certain probability limit theorems*.

Dudley, R. M. (2002). *Real analysis and probability* (2nd ed.). Cambridge University Press. doi: 10.1017/CBO9780511755347

DuMouchel, W. H., & Duncan, G. J. (1983). Using sample survey weights in multiple regression analyses of stratified samples. *Journal of the American Statistical Association*, *78*, 535–543.

Fay, R. E., & Herriot, R. A. (1979). Estimation of income from small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, *74*, 269–277.

Fowler, F. J., Jr., Roman, A. M., & Xiao Di, Z. (1998, 05). Mode Effects in a Survey of Medicare Prostate Surgery Patients . *Public Opinion Quarterly*, *62*(1), 29–46. Retrieved from `https://doi.org/10.1086/297829` doi: 10.1086/297829

Francisco, C. A., & Fuller, W. A. (1991). Quantile estimation with a complex survey design. *Annals of Statistics*, *19*, 454–469.

Fukumizu, K., Gretton, A., Sun, X., & Schölkopf, B. (2008). Kernel measures of conditional dependence. In J. C. Platt, D. Koller, Y. Singer, & S. T. Roweis (Eds.), *Advances in neural information processing systems 20* (pp. 489–496). Curran Associates, Inc. Retrieved from `http://papers.nips.cc/paper/3340-kernel-measures-of-conditional-dependence.pdf`

Fuller, W. A. (1984). Least squares and related analyses for complex survey designs. *Survey Methodology*, *10*, 97–118.

Fuller, W. A. (2009). *Sampling statistics*. Hoboken, NJ: Wiley & Sons.

Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., & Smola, A. J. (2007). A kernel method for the two-sample-problem. In B. Schölkopf, J. C. Platt, & T. Hoffman (Eds.), *Advances in neural information processing systems 19* (pp. 513–520). MIT Press. Retrieved from `http://papers.nips.cc/paper/3110-a-kernel-method-for-the-two-sample-problem.pdf`

Gretton, A., Borgwardt, K., Rasch, M. J., Scholkopf, B., & Smola, A. J. (2008). A kernel method for the two-sample problem. *arXiv preprint arXiv:0805.2368*.

Gretton, A., Borgwardt, K., Rasch, M. J., Scholkopf, B., & Smola, A. J. (2012). A kernel two-sample test. *Journal of Machine Learning Research*, *13*, 723–773.

Gretton, A., Fukumizu, K., Harchaoui, Z., & Sriperumbudur, B. K. (2009). A fast, consistent kernel two-sample test. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, & A. Culotta (Eds.), *Advances in neural information processing systems 22* (pp. 673–681). Curran Associates, Inc. Retrieved from `http://papers.nips.cc/paper/3738-a-fast-consistent-kernel-two-sample-test.pdf`

Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Annals of Mathematical Statistics*, *35*, 1491–1523.

Harville, D. (1985). Decomposition of prediction error. *Journal of the American Statistical Association*, *80*, 132–138.

Herndon, W. (2014). *Testing and adjusting for informative sampling in survey data.* Retrieved from `http://search.proquest.com/docview/1615129365/`

Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, *47*, 663-685.

Jiang, J., & Lahiri, P. (2006). Mixed model prediction and small area estimation. *TEST*, *15*(1), 1. Retrieved from `https://doi.org/10.1007/BF02595419` doi: 10.1007/BF02595419

Kackar, R. N., & Harville, D. A. (1984). Approximations for standard errors of estimators of fixed and random effects in mixed linear models. *Journal of the American Statistical Association*, *79*, 853–862.

Kim, H., Sun, D., & Tsutakawa, R. K. (2001). A bivariate bayes method for improving the estimates of mortality rates with a twofold conditional autoregressive model. *Journal of the American Statistical Association*, *96*(456), 1506-1521. doi: 10.1198/016214501753382408

Kolmogorov, A. (1933). Sulla determinazione empirica di una legge di distribuzione. *G. Ist. Ital. Attuari.*, *4*, 83–91.

Korn, E. L., & Graubard, B. I. (1999). *Analysis of health surveys*. New York: John Wiley & Sons, Inc.

Krieger, A. M., & Pfeffermann, D. (1992). Maximum likelihood estimation from complex sample surveys. *Survey Methodology*, *18*, 225–239.

Molina, I., & Marhuenda, Y. (2015). `sae`: An `R` package for small area estimation. *The R Journal*, *7/1*, 81–98.

Müller, A. (1997). Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, *29*(2), 429–443. Retrieved from `http://www.jstor.org/stable/1428011`

National Research Council. (2006). *Review of recreational fisheries survey methods*. Washington, DC: The National Academies Press.

Nordberg, L. (1989). Generalized linear modeling of sample survey data. *Journal of Official Statistics*, *5*, 223–239.

Opsomer, J., Botts, C., & Kim, J. (2003, 06). Small area estimation in a watershed erosion assessment survey. *Journal of Agricultural, Biological, and Environmental Statistics*, *8*, 139-152. doi: 10.1198/1085711031607

Patil, G. P., & Rao, C. R. (1978). Weighted distributions and size-biased sampling with applications to wildlife populations and human families. *Biometrics*, *34*(2), 179–189. Retrieved from `http://www.jstor.org/stable/2530008`

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review*, *61*, 317–337.

Pfeffermann, D. (2002). Small area estimation: New developments and directions. *International Statistical Review / Revue Internationale de Statistique*, *70*(1), 125–143. Retrieved from http://www.jstor.org/stable/1403729

Pfeffermann, D. (2013, 02). New important developments in small area estimation. *Statist. Sci.*, *28*(1), 40–68. Retrieved from https://doi.org/10.1214/12-STS395 doi: 10.1214/12-STS395

Pfeffermann, D., Da Silva Moura, F. A., & Do Nascimento Silva, P. L. (2006). Multi-level modelling under informative sampling. *Biometrika*, *93*(4), 943-959. Retrieved from http://dx.doi.org/10.1093/biomet/93.4.943 doi: 10.1093/biomet/93.4.943

Pfeffermann, D., & Sverchkov, M. (1999). Parametric and semi-parametric estimation of regression models fitted to survey data. *Sankhyā, Series B*, *61*, 166–186.

Pfeffermann, D., & Sverchkov, M. (2003). Fitting generalized linear models under informative sampling. In *Analysis of survey data* (p. 175-195). Wiley-Blackwell. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1002/0470867205.ch12 doi: 10.1002/0470867205.ch12

Rao, J., & Molina, I. (2015). *Small area estimation*. Wiley. Retrieved from https://books.google.com/books?id=i1B_BwAAQBAJ

Rubin-Bleuer, S., & Kratina, I. S. (2005). On the two-phase framework for joint model and design-based inference. *Annals of Statistics*, *33*, 2789–2810.

Särndal, C. E., Swensson, B., & Wretman, J. (1992). *Model assisted survey sampling*. New York: Springer-Verlag.

Serfling, R. J. (1980). *Approximation theorems of mathematical statistics*. New York: John Wiley & Sons.

Skinner, C. J. (1994). Sample models and weights. In *Asa proceedings of the section on survey research methods* (pp. 133–142). American Statistical Association.

Smirnov, N. V. (1937). On the distribution of the ω^2 criterion of von Mises. *Rec. Math.*, *2*, 973-993.

Smirnov, N. V. (1939). On the estimation of the discrepancy between empirical curves of distribution for two independent sample. *Bull. Math. Univ. Moscow*, *2*(2), 3-14.

Smola, A., Gretton, A., Song, L., & Schölkopf, B. (2007). A hilbert space embedding for distributions. In M. Hutter, R. A. Servedio, & E. Takimoto (Eds.), *Algorithmic learning theory* (pp. 13–31). Berlin, Heidelberg: Springer Berlin Heidelberg.

Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Schölkopf, B., & Lanckriet, G. R. (2010, August). Hilbert space embeddings and metrics on probability measures. *J. Mach. Learn. Res.*, *11*, 1517–1561.

Steinwart, I. (2002, March). On the influence of the kernel on the consistency of support vector machines. *J. Mach. Learn. Res.*, *2*, 67–93. Retrieved from https://doi.org/10.1162/153244302760185252 doi: 10.1162/153244302760185252

Sverchkov, M., & Pfeffermann, D. (2004). Prediction of finite population totals based on the sample distribution. *Survey Methodology*, *30*(1), 79–92.

Thompson, M. (2013). *Theory of sample surveys*. Springer US. Retrieved from https://books.google.com/books?id=n2ezDAEACAAJ

Turner, C., Ku, L., Rogers, S., Lindberg, L., Pleck, J., & Sonenstein, F. (1998, 5 8). Adolescent sexual behavior, drug use, and violence: Increased reporting with computer survey technology. *Science*, *280*(5365), 867–873. doi: 10.1126/science.280.5365.867

US Census Bureau. (2016). State population totals datasets: 2010-2016 [Computer software manual]. (https://www.census.gov/data/datasets/2016/demo/popest/state-total.html)

Van Den Brakel, J., Zhang, X. M., & Tam, S.-M. (2020). Measuring discontinuities in time series obtained with repeated sample surveys. *International Statistical Review*, *n/a*(n/a). Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/insr.12347 doi: 10.1111/insr.12347

Van Den Brakel, J. A. (2008). Design-based analysis of embedded experiments with applications in the dutch labour force survey. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *171*(3), 581-613. doi: 10.1111/j.1467-985X.2008.00532.x

Van Den Brakel, J. A. (2010, 09). Sampling and estimation techniques for the implementation of new classification systems: The change-over from nace rev. 1.1 to nace rev. 2 in business surveys. *Survey Research Methods*, *4*, 103-119. doi: 10.18148/srm/2010.v4i2.2354

Van Den Brakel, J. A. (2013, 12). Design-based analysis of factorial designs embedded in probability samples. *Survey Methodology*, *39*, 323-349.

van der Vaart, A. W., & Wellner, J. A. (1996). *Weak convergence and empirical processes*. Springer-Verlag Inc.

Wolter, K. M. (2007). *Introduction to variance estimation* (2nd ed.). New York: Springer-Verlag Inc.

Appendix A

A.1 Proof of Lemmas and Theorem in Chapter 2

To prove the convergence of $\mathbb{T}_N = \sqrt{n}\left(\widehat{F}_{HJ} - \widehat{F}\right)$ in main theorem, we prove convergence of several intermediate random processes. In Appendix A.1.2, we show $\mathbb{T}_N(t) = \mathbb{G}_N(t) + o_p(1)$, where $\mathbb{G}_N(t) = \frac{\sqrt{n}}{N}\sum_{i=1}^{N}\frac{\xi_i}{\pi_i}\left(1 - \frac{N\pi_i}{n}\right)\left(\mathbb{1}_{(Y_i \leq t)} - F(t)\right)$. Boistard et al. (2017a) show the convergence of process $\frac{\sqrt{n}}{N}\sum_{i=1}^{N}\left(\frac{\xi_i}{\pi_i} - 1\right)\left(\mathbb{1}_{(Y_i \leq t)} - F(t)\right)$. Similarly, we show the convergence of $\mathbb{Y}_N(t) = \frac{\sqrt{n}}{N}\sum_{i=1}^{N}\left(\frac{\xi_i}{\pi_i} - 1\right)\left(1 - \frac{N\pi_i}{n}\right)\left(\mathbb{1}_{(Y_i \leq t)} - F(t)\right)$ in Lemma A.1.3. We see that $\mathbb{Y}_N(t)$ is equivalent to $\mathbb{X}_N(t) - \mathbb{F}_N(t)$, where $\mathbb{X}_N(t) = \frac{\sqrt{n}}{N}\sum_{i=1}^{N}\left(\frac{\xi_i}{\pi_i} - 1\right)\left(1 - \frac{N\pi_i}{n}\right)\mathbb{1}_{(Y_i \leq t)}$ and $\mathbb{F}_N(t) = \frac{\sqrt{n}}{N}\sum_{i=1}^{N}\left(\frac{\xi_i}{\pi_i} - 1\right)\left(1 - \frac{N\pi_i}{n}\right)F(t)$. Both of these two processes have the form of CLT assumption in (H1), so the convergence follows by showing their weak convergence of all finite dimensional distributions and tightness, in Lemma A.1.1 and Lemma A.1.2 respectively.

A.1.1 Lemmas and Proofs

Lemma A.1.1. *Let $Y_1, \ldots, Y_N$ be iid random variables with cdf F. Suppose that conditions (C1)–(C4) and (H1)–(H2) hold, then $\frac{\sqrt{n}}{N}\sum_{i=1}^{N}\left(\frac{\xi_i}{\pi_i} - 1\right)\left(1 - \frac{N\pi_i}{n}\right)\mathbb{1}_{(Y_i \leq t)}$ converges weakly to a zero mean Gaussian process $\mathbb{G}^{HT}$ with covariance function*

$$\mathbb{E}_m\mathbb{G}^{HT}(s)\mathbb{G}^{HT}(t) = \lim_{N\to\infty}\frac{n}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\mathbb{E}_m\left[\frac{\pi_{ij} - \pi_i\pi_j}{\pi_i\pi_j}\left(1 - \frac{N\pi_i}{n}\right)\left(1 - \frac{N\pi_j}{n}\right)\mathbb{1}_{(Y_i \leq s)}\mathbb{1}_{(Y_j \leq t)}\right]$$

for $s, t \in \mathbb{R}$.

Proof. Proof of Lemma A.1.1 is very similar to the proof of Theorem 3.1 in Boistard et al. (2017a). We will use Theorem 13.5 from Billingsley (1999), by first checking the tightness condition, and then constructing the weak convergence in all finite dimensional distributions.

To establish the tightness condition, we will use 13.14 from Billingsley (1999). Suppose (C1)–(C4) hold, let $\mathbb{X}_N = \frac{\sqrt{n}}{N}\sum_{i=1}^{N}\left(\frac{\xi_i}{\pi_i} - 1\right)\left(1 - \frac{N\pi_i}{n}\right)\mathbb{1}_{(Y_i \leq t)}$. If there exists a constant $K > 0$ independent of N, such that for any t_1, t_2 and $-\infty < t_1 \leq t \leq t_2 < \infty$,

$$\mathbb{E}_{d,m}\left[(\mathbb{X}_N(t)-\mathbb{X}_N(t_1))^2(\mathbb{X}_N(t_2)-\mathbb{X}_N(t))^2\right] \le K\left(F(t_2)-F(t_1)\right)^2, \tag{A.1.1}$$

then we claim we have tightness.

First we define $p_1 = F(t)-F(t_1),\ p_2 = F(t_2)-F(t),\ A_i = \mathbb{1}_{\{t_1<Y_i\le t\}}$, and $B_i = \mathbb{1}_{\{t<Y_i\le t_2\}}$. Furthermore, let $\alpha_i = \left(\frac{\xi_i-\pi_i}{\pi_i}\right)\left(1-\frac{N\pi_i}{n}\right)A_i$ and $\beta_i = \left(\frac{\xi_i-\pi_i}{\pi_i}\right)\left(1-\frac{N\pi_i}{n}\right)B_i$. Then, as $p_1p_2 \le (F(t_2)-F(t_1))^2$, it suffices to show

$$\frac{1}{N^4}\mathbb{E}_{d,m}\left[n^2\left(\sum_{i=1}^{N}\alpha_i\right)^2\left(\sum_{j=1}^{N}\beta_j\right)^2\right] \le Kp_1p_2. \tag{A.1.2}$$

The expectation on the left side can be decomposed as

$$\begin{aligned}&N^{-4}\sum_{i=1}^{N}\sum_{k=1}^{N}\mathbb{E}_{d,m}[n^2\alpha_i^2\beta_k^2] + N^{-4}\sum_{i=1}^{N}\sum_{j\neq i}\sum_{k=1}^{N}\mathbb{E}_{d,m}[n^2\alpha_i\alpha_j\beta_k^2]\\ &+N^{-4}\sum_{k=1}^{N}\sum_{l\neq k}\sum_{i=1}^{N}\mathbb{E}_{d,m}[n^2\alpha_i^2\beta_k\beta_l] + N^{-4}\sum_{i=1}^{N}\sum_{j\neq i}\sum_{k=1}^{N}\sum_{l\neq k}\mathbb{E}_{d,m}[n^2\alpha_i\alpha_j\beta_k\beta_l].\end{aligned} \tag{A.1.3}$$

By symmetry, the two triple sums can be handled similarly. So we only need to consider three summations.

Since $\mathbb{1}_{\{t_1<Y_i\le t\}}\mathbb{1}_{\{t<Y_i\le t_2\}} = 0$, we will only have non-zero expectations when $\{i,j\}$ and $\{k,l\}$ are disjoint. So

$$\begin{aligned}\frac{1}{N^4}\sum_{i=1}^{N}\sum_{k=1}^{N}\mathbb{E}_{d,m}[n^2\alpha_i^2\beta_k^2] &= \frac{1}{N^4}\sum\sum_{(i,k)\in D_{2,N}}\mathbb{E}_{d,m}[n^2\alpha_i^2\beta_k^2]\\ &= \frac{1}{N^4}\sum\sum_{(i,k)\in D_{2,N}}\mathbb{E}_m\left[n^2\frac{A_iB_k}{\pi_i^2\pi_k^2}\left(1-\frac{N\pi_i}{n}\right)^2\left(1-\frac{N\pi_k}{n}\right)^2\mathbb{E}_d(\xi_i-\pi_i)^2(\xi_k-\pi_k)^2\right]\\ &\le \frac{1}{K_1^4}\sum\sum_{(i,k)\in D_{2,N}}\mathbb{E}_m\left[\frac{A_iB_k}{n^2}\left(1-\frac{N\pi_i}{n}\right)^2\left(1-\frac{N\pi_k}{n}\right)^2\mathbb{E}_d(\xi_i-\pi_i)^2(\xi_k-\pi_k)^2\right]\\ &\le \frac{1}{K_4^4}\sum\sum_{(i,k)\in D_{2,N}}\mathbb{E}_m\left[\frac{A_iB_k}{n^2}\mathbb{E}_d(\xi_i-\pi_i)^2(\xi_k-\pi_k)^2\right]\end{aligned}$$

for some constant K_4 as a direct result of condition (C1). Straightforward computation shows that $\mathbb{E}_d(\xi_i - \pi_i)^2(\xi_k - \pi_k)^2$ equals

$$(\pi_{ik} - \pi_i\pi_k)(1 - 2\pi_i)(1 - 2\pi_k) + \pi_i\pi_k(1 - \pi_i)(1 - \pi_k).$$

Hence, with (C1)–(C2) we find that

$$\mathbb{E}_d(\xi_i - \pi_i)^2(\xi_k - \pi_k)^2 \leq |\mathbb{E}_d(\xi_i - \pi_i)(\xi_k - \pi_k)| + K_2^2\frac{n^2}{N^2} = \mathcal{O}\left(\frac{n^2}{N^2}\right),$$

ω-a.s. It follows that

$$\frac{1}{N^4}\sum_{i=1}^{N}\sum_{k=1}^{N}\mathbb{E}_{d,m}[n^2\alpha_i^2\beta_k^2] \leq \mathcal{O}\left(\frac{1}{N^2}\right)\sum\sum_{(i,k)\in D_{2,N}}\mathbb{E}_m[A_iB_k].$$

Since $D_{2,N}$ has $\mathcal{O}\left(N^2\right)$ elements and $\mathbb{E}_m[A_iB_j] = p_1p_2$ for $(i,j) \in D_{2,N}$, it follows that

$$\frac{1}{N^4}\sum_{i=1}^{N}\sum_{k=1}^{N}\mathbb{E}_{d,m}[n^2\alpha_i^2\beta_k^2] \leq Kp_1p_2. \tag{A.1.4}$$

The second summation of (A.1.3) can be written as:

$$\begin{aligned}
\frac{1}{N^4}\left|\sum_{i=1}^{N}\sum_{j\neq i}\sum_{k=1}^{N}\mathbb{E}_{d,m}[n^2\alpha_i\alpha_j\beta_k^2]\right| &= \frac{1}{N^4}\left|\sum\sum\sum_{(i,j,k)\in D_{3,N}}\mathbb{E}_{d,m}[n^2\alpha_i\alpha_j\beta_k^2]\right| \\
&\leq \frac{1}{N^4}\sum\sum\sum_{(i,j,k)\in D_{3,N}}\mathbb{E}_m\left[n^2\frac{A_iA_jB_k}{\pi_i\pi_j\pi_k^2}\left(1-\frac{N\pi_i}{n}\right)\left(1-\frac{N\pi_j}{n}\right)\left(1-\frac{N\pi_k}{n}\right)^2\right. \\
&\qquad \left.\times\left|\mathbb{E}_d(\xi_i - \pi_i)(\xi_j - \pi_j)(\xi_k - \pi_k)^2\right|\right] \\
&\leq \frac{1}{K_1^4}\sum\sum\sum_{(i,j,k)\in D_{3,N}}\mathbb{E}_m\left[\frac{A_iA_jB_k}{n^2}\left(1-\frac{N\pi_i}{n}\right)\left(1-\frac{N\pi_j}{n}\right)\left(1-\frac{N\pi_k}{n}\right)^2\right. \\
&\qquad \left.\times\left|\mathbb{E}_d(\xi_i - \pi_i)(\xi_j - \pi_j)(\xi_k - \pi_k)^2\right|\right] \\
&\leq \frac{1}{K_4^4}\sum\sum\sum_{(i,j,k)\in D_{3,N}}\mathbb{E}_m\left[\frac{A_iA_jB_k}{n^2}\left|\mathbb{E}_d(\xi_i - \pi_i)(\xi_j - \pi_j)(\xi_k - \pi_k)^2\right|\right].
\end{aligned}$$

We find that $\mathbb{E}_d(\xi_i - \pi_i)(\xi_j - \pi_j)(\xi_k - \pi_k)^2$ equals

$$(1-2\pi_k)\mathbb{E}_d(\xi_i - \pi_i)(\xi_j - \pi_j)(\xi_k - \pi_k) + \pi_k(1-\pi_k)\mathbb{E}_d(\xi_i - \pi_i)(\xi_j - \pi_j).$$

With (C1)–(C3), this means $|\mathbb{E}_d(\xi_i - \pi_i)(\xi_j - \pi_j)(\xi_k - \pi_k)^2| = \mathcal{O}\left(n^2/N^3\right)$, ω-a.s. It follows that

$$\frac{1}{N^4}\left|\sum_{i=1}^{N}\sum_{j\neq i}\sum_{k=1}^{N}\mathbb{E}_{d,m}[n^2\alpha_i\alpha_j\beta_k^2]\right| = \mathcal{O}\left(\frac{1}{N^3}\right)\sum\sum\sum_{(i,j,k)\in D_{3,N}}\mathbb{E}_m[A_iA_jB_k].$$

Since $D_{3,N}$ has $\mathcal{O}\left(N^3\right)$ elements and $\mathbb{E}_m[A_iA_jB_k] = p_1^2p_2$, for $(i,j,k)\in D_{3,N}$, we find

$$\frac{1}{N^4}\left|\sum_{i=1}^{N}\sum_{j\neq i}\sum_{k=1}^{N}\mathbb{E}_{d,m}[n^2\alpha_i\alpha_j\beta_k^2]\right| \leq Kp_1p_2. \tag{A.1.5}$$

The third summation in (A.1.3) is bounded by the same argument leading to (A.1.5). Finally we consider the last summation in (A.1.3):

$$\begin{aligned}
&\frac{1}{N^4}\left|\sum_{i=1}^{N}\sum_{j\neq i}\sum_{k=1}^{N}\sum_{l\neq k}\mathbb{E}_{d,m}[n^2\alpha_i\alpha_j\beta_k\beta_l]\right| = \frac{1}{N^4}\left|\sum\sum\sum\sum_{(i,j,k,l)\in D_{4,N}}\mathbb{E}_{d,m}[n^2\alpha_i\alpha_j\beta_k\beta_l]\right| \\
&\leq \frac{1}{K_1^4}\sum\sum\sum\sum_{(i,j,k,l)\in D_{4,N}}\mathbb{E}_m\left[\frac{A_iA_jB_kB_l}{n^2}\left(1-\frac{N\pi_i}{n}\right)\left(1-\frac{N\pi_j}{n}\right)\right. \\
&\qquad \left.\times\left(1-\frac{N\pi_k}{n}\right)\left(1-\frac{N\pi_l}{n}\right)|\mathbb{E}_d(\xi_i-\pi_i)(\xi_j-\pi_j)(\xi_k-\pi_k)(\xi_l-\pi_l)|\right] \\
&\leq \frac{1}{K_4^4}\sum\sum\sum\sum_{(i,j,k,l)\in D_{4,N}}\mathbb{E}_m\left[\frac{A_iA_jB_kB_l}{n^2}|\mathbb{E}_d(\xi_i-\pi_i)(\xi_j-\pi_j)(\xi_k-\pi_k)(\xi_l-\pi_l)|\right].
\end{aligned}$$

Since $D_{4,N}$ has $\mathcal{O}\left(N^4\right)$ elements and $\mathbb{E}_m\left[A_iA_jB_kB_l\right] = p_1^2p_2^2$, for $(i,j,k,l)\in D_{4,N}$, with (C4), we conclude that

$$\frac{1}{N^4}\left|\sum_{i=1}^{N}\sum_{j\neq i}\sum_{k=1}^{N}\sum_{l\neq k}\mathbb{E}_{d,m}[n^2\alpha_i\alpha_j\beta_k\beta_l]\right| \leq Kp_1p_2. \tag{A.1.6}$$

Combining (A.1.4), (A.1.5), (A.1.6) and (A.1.3), we get (A.1.1), concluding the proof of tightness. Next we prove all finite dimensional weak convergence by Cramér-Wold device.

As $\mathbb{X}_N = \frac{\sqrt{n}}{N} \sum_{i=1}^N \left(\frac{\xi_i}{\pi_i} - 1\right) \left(1 - \frac{N\pi_i}{n}\right) \mathbb{1}_{(Y_i \leq t)}$, then

$$a_1 \mathbb{X}_N(t_1) + \cdots + a_k \mathbb{X}_N(t_k) = \frac{\sqrt{n}}{N} \sum_{i=1}^N \left(\frac{\xi_i}{\pi_i} - 1\right) V_{ik}, \tag{A.1.7}$$

where

$$V_{ik} = a_1 \left(1 - \frac{N\pi_i}{n}\right) \mathbb{1}_{(Y_i \leq t_1)} + \cdots + a_k \left(1 - \frac{N\pi_i}{n}\right) \mathbb{1}_{(Y_i \leq t_k)} = \mathbf{a}_k^\intercal \mathbf{Y}_{ik}^*, \tag{A.1.8}$$

with $\mathbf{Y}_{ik}^{*\intercal} = \left(1 - \frac{N\pi_i}{n}\right) \left(\mathbb{1}_{(Y_i \leq t_1)}, \ldots, \mathbb{1}_{(Y_i \leq t_k)}\right)$ and $\mathbf{a}_k^\intercal = (a_1, \ldots, a_k)$. For the design-based variance, we have

$$\begin{aligned} nS_N^2 &= \frac{n}{N^2} \sum_{i=1}^N \sum_{j=1}^N \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} V_{ik} V_{jk} && \text{(A.1.9)} \\ &= \mathbf{a}_k^\intercal \left(\frac{n}{N^2} \sum_{i=1}^N \sum_{j=1}^N \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \mathbf{Y}_{ik}^* \mathbf{Y}_{jk}^{*\intercal} \right) \mathbf{a}_k \to \mathbf{a}_k^\intercal \boldsymbol{\Sigma}_k^* \mathbf{a}_k, && \text{(A.1.10)} \end{aligned}$$

according to (H2). Together with (H1), it follows that (A.1.7) converges in distribution to a zero mean normal random variable with variance $\mathbf{a}_k^\intercal \boldsymbol{\Sigma}_k^* \mathbf{a}_k$. We conclude that $(\mathbb{X}_N(t_1), \ldots, \mathbb{X}_N(t_k))$ has a k-variate zero mean normal distribution with covariance matrix $\boldsymbol{\Sigma}_k^*$. According to the Cramér-Wold device, this proves weak convergence of all the finite-dimensional distributions.

Combining with the tightness condition we established previously, we have proved the lemma for the case that Y_i's are uniformly distributed on $[0, 1]$. Extension of this to Y_i's with a general cdf F can be found at Boistard et al. (2017a). □

Lemma A.1.2. *Let $Y_1, \ldots, Y_N$ be iid random variables with cdf F. Suppose that conditions (C1)–(C4) and (H1)–(H2) hold, then $\frac{\sqrt{n}}{N} \sum_{i=1}^N \left(\frac{\xi_i}{\pi_i} - 1\right) \left(1 - \frac{N\pi_i}{n}\right) F(t)$ converges weakly to a Gaussian process $\mathbb{G}_F^*$.*

Proof. Set $V_i = 1 - \pi_i N/n$ in (H1), the limiting Gaussian process easily follows. □

Lemma A.1.3. *Let $Y_1, \dots, Y_N$ be iid random variables with cdf F. Suppose that conditions (C1)–(C4), (H1) and (H3) hold, then*

$$\mathbb{Y}_N(t) = \frac{\sqrt{n}}{N} \sum_{i=1}^{N} \left(\frac{\xi_i}{\pi_i} - 1 \right) \left(1 - \frac{N\pi_i}{n} \right) \left(\mathbb{1}_{(Y_i \le t)} - F(t) \right)$$

converges weakly to a mean zero Gaussian process $\mathbb{G}_F^{HT}$ with covariance function

$$\mathbb{E}_m \mathbb{G}_F^{HT}(s) \mathbb{G}_F^{HT}(t) =$$
$$\lim_{N\to\infty} \frac{n}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \mathbb{E}_m \left[\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \left(1 - \frac{N\pi_i}{n} \right) \left(1 - \frac{N\pi_j}{n} \right) \left(\mathbb{1}_{(Y_i \le s)} - F(s) \right) \left(\mathbb{1}_{(Y_j \le t)} - F(t) \right) \right]$$

for $s, t \in \mathbb{R}$.

Proof. Because $\mathbb{Y}_N(t)$ is the difference between the two random quantities in Lemma A.1.1 and Lemma A.1.2, each of which converges to a tight continuous process, the tightness of $\mathbb{Y}_N(t)$ follows from Lemma $B.2$ in Boistard et al. (2017b).

Convergence of the finite dimensional distributions is similar to Lemma A.1.1. As $\mathbb{Y}_N(t) = \frac{\sqrt{n}}{N} \sum_{i=1}^{N} \left(\frac{\xi_i}{\pi_i} - 1 \right) \left(1 - \frac{N\pi_i}{n} \right) \left(\mathbb{1}_{(Y_i \le t)} - F(t) \right)$, then

$$a_1 Y_N(t_1) + \cdots + a_k Y_N(t_k) = \frac{\sqrt{n}}{N} \sum_{i=1}^{N} \left(\frac{\xi_i}{\pi_i} - 1 \right) V_{ik}, \tag{A.1.11}$$

where

$$V_{ik} = a_1 \left(1 - \frac{N\pi_i}{n} \right) \left(\mathbb{1}_{(Y_i \le t_1)} - F(t_1) \right) + \cdots + a_k \left(1 - \frac{N\pi_i}{n} \right) \left(\mathbb{1}_{(Y_i \le t_K)} - F(t_k) \right) \tag{A.1.12}$$
$$= \mathbf{a}_k^{\intercal} \mathbf{Y}_{ik},$$

with $\mathbf{Y}_{ik}^{\intercal} = \left(1 - \frac{N\pi_i}{n} \right) \left(\mathbb{1}_{(Y_i \le t_1)} - F(t_1), \dots, \mathbb{1}_{(Y_i \le t_k)} - F(t_k) \right)$ and $\mathbf{a}_k^{\intercal} = (a_1, \dots, a_k)$. For the design-based variance, we have

$$nS_N^2 = \frac{n}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}V_{ik}V_{jk} \tag{A.1.13}$$

$$= \mathbf{a}_k^\intercal\left(\frac{n}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}\mathbf{Y}_{ik}\mathbf{Y}_{jk}^\intercal\right)\mathbf{a}_k \to \mathbf{a}_k^\intercal\mathbf{\Sigma}_k\mathbf{a}_k, \tag{A.1.14}$$

according to (H3). Together with (H1), it follows that (A.1.11) converges in distribution to a zero mean normal random variable with variance $\mathbf{a}_k^\intercal\mathbf{\Sigma}_k\mathbf{a}_k$. We conclude that $(\mathbb{Y}_N(t_1),\dots,\mathbb{Y}_N(t_k))$ has a k-variate zero mean normal distribution with covariance matrix $\mathbf{\Sigma}_k$. According to the Cramér-Wold device, this proves weak convergence of all finite dimensional distributions.

□

Lemma A.1.4. *Let $Y_1,\dots,Y_N$ be iid random variables with cdf F. Suppose that conditions (C1)–(C4), (H1) and (H3)–(H5) hold, then $\mathbb{G}_N(t) = \frac{\sqrt{n}}{N}\sum_{i=1}^{N}\frac{\xi_i}{\pi_i}\left(1-\frac{N\pi_i}{n}\right)\left(\mathbb{1}_{(Y_i\le t)}-F(t)\right)$ converges weakly to a mean zero Gaussian process $\mathbb{G}_F$ with covariance function*

$$\mathbb{E}_m\mathbb{G}_F(s)\mathbb{G}_F(t) =$$
$$\lim_{N\to\infty}\frac{n}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\mathbb{E}_m\left[\frac{\pi_{ij}}{\pi_i\pi_j}\left(1-\frac{N\pi_i}{n}\right)\left(1-\frac{N\pi_j}{n}\right)\left(\mathbb{1}_{(Y_i\le s)}-F(s)\right)\left(\mathbb{1}_{(Y_j\le t)}-F(t)\right)\right]$$

for $s,t\in\mathbb{R}$.

Proof. We first decompose $\mathbb{G}_N(t)$ into two parts,

$$\mathbb{G}_N(t) = \frac{\sqrt{n}}{N}\sum_{i=1}^{N}\frac{\xi_i}{\pi_i}\left(1-\frac{N\pi_i}{n}\right)\left(\mathbb{1}_{(Y_i\le t)}-F(t)\right)$$
$$= \mathbb{Y}_N(t) + \frac{\sqrt{n}}{N}\sum_{i=1}^{N}\left(1-\frac{N\pi_i}{n}\right)\left(\mathbb{1}_{(Y_i\le t)}-F(t)\right)$$
$$= \mathbb{Y}_N(t) + \mathbb{Z}_N(t). \tag{A.1.15}$$

We have shown the limiting process of $\mathbb{Y}_N(t)$ in Lemma A.1.3. The limiting process of $\mathbb{Z}_N(t)$ can be established by similar arguments to Lemma A.1.2. Thus we have the sum of two tight continuous limiting processes, implying that $\mathbb{G}_N(t)$ is tight. Next, we find the covariance function of $\mathbb{G}_N(t)$. We know

$$\text{Var}\left(\mathbb{G}_N(t)\right) = \text{Var}\left(\mathbb{Y}_N(t)\right) + \text{Var}\left(\mathbb{Z}_N(t)\right) + 2\,\text{Cov}\left(\mathbb{Y}_N(t), \mathbb{Z}_N(t)\right), \tag{A.1.16}$$

and

$$\begin{aligned}\text{Cov}\left(\mathbb{Y}_N(t), \mathbb{Z}_N(t)\right) &= \text{Cov}_m\left(\text{E}_\xi\left[\mathbb{Y}_N(t)|\mathbf{Y}, \boldsymbol{\pi}\right], \text{E}_\xi\left[\mathbb{Z}_N(t)|\mathbf{Y}, \boldsymbol{\pi}\right]\right)\\ &\quad + \text{E}_m\left[\text{Cov}_\xi\left(\mathbb{Y}_N(t), \mathbb{Z}_N(t)|\mathbf{Y}, \boldsymbol{\pi}\right)\right]\\ &= \text{Cov}_m\left(0, \mathbb{Z}_N(t)\right) + \text{E}_m\left[0\right] = 0.\end{aligned}$$

It follows that the limiting process has covariance function

$$\begin{aligned}&\mathbb{E}_m \mathbb{G}_F(s)\mathbb{G}_F(t) =\\ &\lim_{N\to\infty} \frac{n}{N^2} \sum_{i=1}^{N}\sum_{j=1}^{N} \mathbb{E}_m\left[\frac{\pi_{ij}}{\pi_i\pi_j}\left(1 - \frac{N\pi_i}{n}\right)\left(1 - \frac{N\pi_j}{n}\right)\left(\mathbb{1}_{(Y_i\le s)} - F(s)\right)\left(\mathbb{1}_{(Y_j\le t)} - F(t)\right)\right]\end{aligned}$$

for $s, t \in \mathbb{R}$. □

A.1.2 Proof of Theorem 2.3.1

Proof. We prove the theorem by showing $\mathbb{T}_N = \mathbb{G}_N + o_p(1)$. Because

$$\frac{\sqrt{n}}{\widehat{N}} \sum_{i=1}^{N} \frac{\xi_i}{\pi_i}\left(1 - \frac{\widehat{N}\pi_i}{n}\right) F(t) = 0,$$

we have

$$\begin{aligned}\mathbb{T}_N &= \sqrt{n}\left(\widehat{F}_{HJ} - \widehat{F}\right) = \sqrt{n}\left(\frac{1}{\widehat{N}}\sum_{i=1}^{N}\frac{\xi_i}{\pi_i}\mathbb{1}_{(Y_i \le t)} - \frac{1}{n}\sum_{i=1}^{N}\xi_i\mathbb{1}_{(Y_i\le t)}\right)\\ &= \frac{\sqrt{n}}{\widehat{N}}\sum_{i=1}^{N}\frac{\xi_i}{\pi_i}\left(1 - \frac{\widehat{N}\pi_i}{n}\right)\left(\mathbb{1}_{(Y_i\le t)} - F(t)\right)\\ &= \frac{\sqrt{n}}{\widehat{N}}\sum_{i=1}^{N}\frac{\xi_i}{\pi_i}\left(\mathbb{1}_{(Y_i\le t)} - F(t)\right) - \frac{1}{\sqrt{n}}\sum_{i=1}^{N}\xi_i\left(\mathbb{1}_{(Y_i\le t)} - F(t)\right)\\ &= \frac{N}{\widehat{N}}\frac{\sqrt{n}}{N}\sum_{i=1}^{N}\frac{\xi_i}{\pi_i}\left(\mathbb{1}_{(Y_i\le t)} - F(t)\right) - \frac{\sqrt{n}}{N}\sum_{i=1}^{N}\frac{\xi_i}{\pi_i}\frac{N\pi_i}{n}\left(\mathbb{1}_{(Y_i\le t)} - F(t)\right)\\ &= \mathbb{G}_N + o_p(1)\end{aligned}$$

since $\widehat{N}/N \xrightarrow{\text{P}} 1$. □

A.2 Proof of Theorem 3.3.1

Proof. Our statistic is

$$n\text{MMD}^2_{\widehat{h}} = \frac{1}{n-1}\sum\sum_{i,j\in s, i\neq j}\left(1 - \frac{w_i}{\overline{w}_s}\right)\left(1 - \frac{w_j}{\overline{w}_s}\right)\widehat{h}(y_i, y_j), \tag{A.2.1}$$

with empirically-centered kernel

$$\widehat{h}(y_i, y_j) = k(y_i, y_j) - \frac{1}{n}\sum_{j\in s}k(y_i, y_j) - \frac{1}{n}\sum_{i\in s}k(y_i, y_j) + \frac{1}{n^2}\sum\sum_{i,j\in s}k(y_i, y_j).$$

Denote

$$n\text{MMD}^2_{h} = \frac{1}{n-1}\sum\sum_{i,j\in s, i\neq j}\left(1 - \frac{w_i}{\overline{w}_s}\right)\left(1 - \frac{w_j}{\overline{w}_s}\right)h(y_i, y_j), \tag{A.2.2}$$

with the theoretically-centered kernel

$$h(y_i, y_j) := k(y_i, y_j) - \mathbb{E}_y k(y_i, y) - \mathbb{E}_y k(y, y_j) + \mathbb{E}_{y,y'}k(y, y'),$$

replacing the empirically-centered kernel $\widehat{h}$. Here, $\mathbb{E}_{y,y'}k(y, y')$ denotes expectation with respect to independent y and y', so that for $i \neq j$,

$$\mathbb{E}_{y,y'}h(y_i, y_j) = 0.$$

We first show $n\mathrm{MMD}^2_{\widehat{h}} - n\mathrm{MMD}^2_{h} \xrightarrow{\mathrm{P}} 0$:

$$\begin{aligned}
& n\mathrm{MMD}^2_{\widehat{h}} - n\mathrm{MMD}^2_{h} \\
&= \frac{1}{n-1}\sum\sum_{i,j\in s, i\neq j}\left(1-\frac{w_i}{\overline{w}_s}\right)\left(1-\frac{w_j}{\overline{w}_s}\right)\left(\widehat{h}(y_i,y_j)-h(y_i,y_j)\right) \\
&= \frac{1}{n-1}\sum\sum_{i,j\in s, i\neq j}\left(1-\frac{w_i}{\overline{w}_s}\right)\left(1-\frac{w_j}{\overline{w}_s}\right)\left(\frac{1}{n^2}\sum\sum_{i',j'\in s}k(y_{i'},y_{j'})-\mathbb{E}_{y,y'}k(y,y')\right) \\
&\quad - \frac{1}{n-1}\sum\sum_{i,j\in s, i\neq j}\left(1-\frac{w_i}{\overline{w}_s}\right)\left(1-\frac{w_j}{\overline{w}_s}\right)\left(\frac{1}{n}\sum_{j'\in s}k(y_i,y_{j'})-\mathbb{E}_{y}k(y_i,y)\right) \\
&\quad - \frac{1}{n-1}\sum\sum_{i,j\in s, i\neq j}\left(1-\frac{w_i}{\overline{w}_s}\right)\left(1-\frac{w_j}{\overline{w}_s}\right)\left(\frac{1}{n}\sum_{i'\in s}k(y_{i'},y_j)-\mathbb{E}_{y}k(y,y_j)\right) \\
&= \mathrm{I}+\mathrm{II}+\mathrm{III}.
\end{aligned}$$

Using $\phi(y)$ to denote $k(y,\cdot)$, we consider the kernel difference in I:

$$\begin{aligned}
& \frac{1}{n^2}\sum\sum_{i',j'\in s}k(y_{i'},y_{j'})-\mathbb{E}_{y,y'}k(y,y') \\
&= \left\langle \frac{1}{n}\sum_{i'\in s}\phi(y_{i'}), \frac{1}{n}\sum_{j'\in s}\phi(y_{j'})\right\rangle - \left\langle \mathbb{E}\phi(y), \mathbb{E}\phi(y)\right\rangle \\
&= \left\langle \frac{1}{n}\sum_{i'\in s}\phi(y_{i'}), \frac{1}{n}\sum_{j'\in s}\phi(y_{j'})\right\rangle - \left\langle \mathbb{E}\phi(y), \frac{1}{n}\sum_{j'\in s}\phi(y_{j'})\right\rangle \\
&\quad + \left\langle \mathbb{E}\phi(y), \frac{1}{n}\sum_{j'\in s}\phi(y_{j'})\right\rangle - \left\langle \mathbb{E}\phi(y), \mathbb{E}\phi(y)\right\rangle \\
&= \left\langle \frac{1}{n}\sum_{i'\in s}\phi(y_{i'})-\mathbb{E}\phi(y), \frac{1}{n}\sum_{j'\in s}\phi(y_{j'})\right\rangle + \left\langle \mathbb{E}\phi(y), \frac{1}{n}\sum_{j'\in s}\phi(y_{j'})-\mathbb{E}\phi(y)\right\rangle \\
&= \mathcal{O}_p\left(n^{-1/2}\right)
\end{aligned}$$

by the central limit theorem and continuity of inner product, and

$$\frac{1}{n-1}\sum\sum_{i,j\in s, i\neq j}\left(1-\frac{w_i}{\overline{w}_s}\right)\left(1-\frac{w_j}{\overline{w}_s}\right)=\mathcal{O}_p\left(1\right)$$

by the central limit theorem. Hence, $\mathrm{I}=\mathcal{O}_p\left(n^{-1/2}\right)$. Because II and III are symmetric, we only consider II. The kernel difference in II is

$$\begin{aligned}\frac{1}{n}\sum_{j'\in s}k(y_i,y_{j'})-\mathbb{E}_y k(y_i,y)&=\left\langle\phi(y_i),\frac{1}{n}\sum_{j'\in s}\phi(y_{j'})\right\rangle-\left\langle\phi(y_i),\mathbb{E}\phi(y)\right\rangle\\&=\left\langle\phi(y_i),\frac{1}{n}\sum_{j'\in s}\phi(y_{j'})-\mathbb{E}\phi(y)\right\rangle,\end{aligned}$$

so that

$$\begin{aligned}\mathrm{II}&=\frac{1}{n-1}\sum\sum_{i,j\in s,i\neq j}\left(1-\frac{w_i}{\overline{w}_s}\right)\left(1-\frac{w_j}{\overline{w}_s}\right)\left(\frac{1}{n}\sum_{j'\in s}k(y_i,y_{j'})-\mathbb{E}_y k(y_i,y)\right)\\&=\frac{1}{n-1}\sum\sum_{i,j\in s}\left(1-\frac{w_i}{\overline{w}_s}\right)\left(1-\frac{w_j}{\overline{w}_s}\right)\left\langle\phi(y_i),\frac{1}{n}\sum_{j'\in s}\phi(y_{j'})-\mathbb{E}\phi(y)\right\rangle\\&\quad-\frac{1}{n-1}\sum_{i\in s}\left(1-\frac{w_i}{\overline{w}_s}\right)^2\left\langle\phi(y_i),\frac{1}{n}\sum_{j'\in s}\phi(y_{j'})-\mathbb{E}\phi(y)\right\rangle\\&=\frac{n}{n-1}\frac{1}{\sqrt{n}}\sum_{j\in s}\left(1-\frac{w_j}{\overline{w}_s}\right)\left\langle\frac{1}{\sqrt{n}}\sum_{i\in s}\left(1-\frac{w_i}{\overline{w}_s}\right)\phi(y_i),\frac{1}{n}\sum_{j'\in s}\phi(y_{j'})-\mathbb{E}\phi(y)\right\rangle\\&\quad-\frac{n}{n-1}\left\langle\frac{1}{n}\sum_{i\in s}\left(1-\frac{w_i}{\overline{w}_s}\right)^2\phi(y_i),\frac{1}{n}\sum_{j'\in s}\phi(y_{j'})-\mathbb{E}\phi(y)\right\rangle\\&=\mathcal{O}_p\left(n^{-1/2}\right),\end{aligned}$$

by the central limit theorem and continuity of inner product.

Hence $n\mathrm{MMD}^2_{\hat{h}}-n\mathrm{MMD}^2_h\xrightarrow{\mathrm{P}}0$, and we can work with $n\mathrm{MMD}^2_h$ to get the limiting distribution under the null.

We now follow Gretton et al. (2008) and Serfling (1980), section 5.5.2, by computing the eigendecomposition of the kernel with respect to the common null probability measure $p=q$:

$$h(y_i,y_j)=\sum_{l=1}^{\infty}\lambda_l\psi_l(y_i)\psi_l(y_j),$$

where λ_l are eigenvalues and $\psi_l(\cdot)$ are orthonormal eigenfunctions in the sense that

$$\mathbb{E}\psi_l(y)\psi_{l'}(y) = \int_{\mathscr{Y}} \psi_l(y)\psi_{l'}(y)\,dp(y) = \begin{cases} 0, & l \neq l' \\ 1, & l = l'. \end{cases}$$

Further, $\mathbb{E}\psi_l(y) = 0$ for all l (see equation (29) of Gretton et al. (2008)), so that Var $(\psi_l(y)) = 1$. Using the eigendecomposition, we have

$$\begin{aligned} n\mathrm{MMD}_h^2 &= \frac{1}{n-1}\sum\sum_{i,j\in s, i\neq j}\left(1-\frac{w_i}{\overline{w}_s}\right)\left(1-\frac{w_j}{\overline{w}_s}\right)h(y_i,y_j) \\ &= \frac{1}{n-1}\sum_l \lambda_l \sum\sum_{i,j\in s, i\neq j}\left(1-\frac{w_i}{\overline{w}_s}\right)\left(1-\frac{w_j}{\overline{w}_s}\right)\psi_l(y_i)\psi_l(y_j) \\ &= \frac{1}{n-1}\sum_l \lambda_l \left[\left(\sum_{i\in s}\left(1-\frac{w_i}{\overline{w}_s}\right)\psi_l(y_i)\right)^2 - \sum_{i\in s}\left(1-\frac{w_i}{\overline{w}_s}\right)^2\psi_l^2(y_i)\right] \\ &= \frac{n}{n-1}\sum_l \lambda_l \left[\left(\frac{1}{\sqrt{n}}\sum_{i\in s}\left(1-\frac{w_i}{\overline{w}_s}\right)\psi_l(y_i)\right)^2 - \frac{1}{n}\sum_{i\in s}\left(1-\frac{w_i}{\overline{w}_s}\right)^2\psi_l^2(y_i)\right]. \end{aligned}$$

Since w_i and y_i are independent under the null hypothesis of noninformativeness,

$$\frac{1}{\sqrt{n}}\sum_{i\in s}\left(1-\frac{w_i}{\overline{w}_s}\right)\psi_l(y_i) \xrightarrow{\mathcal{L}} z_l \sim \mathcal{N}\left(0, \frac{\sigma_w^2}{\mu_w^2}\right),$$

independent across $l = 1, 2, \ldots$, and

$$\frac{1}{n}\sum_{i\in s}\left(1-\frac{w_i}{\overline{w}_s}\right)^2\psi_l^2(y_i) \xrightarrow{\mathrm{P}} \frac{\sigma_2^2}{\mu_w^2}.$$

Hence

$$n\mathrm{MMD}_h^2 \xrightarrow{\mathcal{L}} \sum_{l=1}^{\infty}\lambda_l\left(z_l^2 - \frac{\sigma_w^2}{\mu_w^2}\right), \tag{A.2.3}$$

where $\{z_l\}$ are iid $\mathcal{N}(0, \sigma_w^2/\mu_w^2)$. □

A.3 Proof of MSE Approximation in Section 4.3.3

Proof. By giving and taking, we have

$$\begin{aligned}\phi_{st}\left(\widehat{\boldsymbol{\beta}},\hat{\psi}\right)-\phi_{st}=\left[\phi_{st}\left(\boldsymbol{\beta},\psi\right)-\phi_{st}\right]+\left[\phi_{st}\left(\widetilde{\boldsymbol{\beta}}_{\psi},\psi\right)-\phi_{st}\left(\boldsymbol{\beta},\psi\right)\right]\\+\left[\phi_{st}\left(\widehat{\boldsymbol{\beta}},\hat{\psi}\right)-\phi_{st}\left(\widetilde{\boldsymbol{\beta}}_{\psi},\psi\right)\right].\end{aligned}$$

For some arbitrary function $h(\cdot)$,

$$\mathrm{Cov}\left(h(Y_{st}),\phi_{st}\left(\boldsymbol{\beta},\psi\right)-\phi_{st}\right)=\mathrm{Cov}\left(h(Y_{st}),\nu_{st}\left(\boldsymbol{\beta},\psi\right)-\nu_{st}\right)=0,$$

where the second equality is by (3.10) in Harville (1985). So we have

$$\mathrm{Cov}\left(\phi_{st}\left(\widetilde{\boldsymbol{\beta}}_{\psi},\psi\right)-\phi_{st}\left(\boldsymbol{\beta},\psi\right),\phi_{st}\left(\boldsymbol{\beta},\psi\right)-\phi_{st}\right)=0,\tag{A.3.1}$$

and

$$\mathrm{Cov}\left(\phi_{st}\left(\widehat{\boldsymbol{\beta}},\hat{\psi}\right)-\phi_{st}\left(\widetilde{\boldsymbol{\beta}}_{\psi},\psi\right),\phi_{st}\left(\boldsymbol{\beta},\psi\right)-\phi_{st}\right)=0.\tag{A.3.2}$$

Next we will show

$$\mathrm{Cov}\left(\phi_{st}\left(\widehat{\boldsymbol{\beta}},\hat{\psi}\right)-\phi_{st}\left(\widetilde{\boldsymbol{\beta}}_{\psi},\psi\right),\phi_{st}\left(\widetilde{\boldsymbol{\beta}}_{\psi},\psi\right)-\phi_{st}\left(\boldsymbol{\beta},\psi\right)\right)=0.\tag{A.3.3}$$

From section 2.1 in Kackar & Harville (1984), we know

$$
\begin{aligned}
&\operatorname{Cov}\left(\mathbf{L}^{\intercal}\mathbf{Y}, \phi_{st}\left(\widetilde{\boldsymbol{\beta}}_{\psi}, \psi\right) - \phi_{st}\right) \\
&= \operatorname{Cov}\left(\mathbf{L}^{\intercal}\mathbf{Y}, \phi_{st}\left(\widetilde{\boldsymbol{\beta}}_{\psi}, \psi\right) - \nu_{st}\right) \\
&= \operatorname{Cov}\left(\mathbf{L}^{\intercal}\mathbf{Y}, \nu_{st}\left(\widetilde{\boldsymbol{\beta}}_{\psi}, \psi\right) - \nu_{st}\right) + \operatorname{Cov}\left(\mathbf{L}^{\intercal}\mathbf{Y}, (\mathbf{z}_{st}^{\intercal} - \mathbf{x}_{st}^{\intercal})\widetilde{\boldsymbol{\beta}}(\psi)\right) \\
&= \operatorname{Cov}\left(\mathbf{L}^{\intercal}\mathbf{Y}, \nu_{st}\left(\widetilde{\boldsymbol{\beta}}_{\psi}, \psi\right) - \nu_{st}\right) \\
&\qquad + \operatorname{Cov}\left(\mathbf{L}^{\intercal}\mathbf{Y}, (\mathbf{z}_{st}^{\intercal} - \mathbf{x}_{st}^{\intercal})(\mathbf{X}^{\intercal}\boldsymbol{\Sigma}^{-1}(\psi)\mathbf{X})^{-1}\mathbf{X}^{\intercal}\boldsymbol{\Sigma}^{-1}(\psi)\mathbf{Y}\right) \\
&= \operatorname{Cov}\left(\mathbf{L}^{\intercal}\mathbf{Y}, \nu_{st}\left(\widetilde{\boldsymbol{\beta}}_{\psi}, \psi\right) - \nu_{st}\right) + (\mathbf{z}_{st}^{\intercal} - \mathbf{x}_{st}^{\intercal})(\mathbf{X}^{\intercal}\boldsymbol{\Sigma}^{-1}(\psi)\mathbf{X})^{-1}\mathbf{X}^{\intercal}\mathbf{L} \\
&= \operatorname{Cov}\left(\mathbf{L}^{\intercal}\mathbf{Y}, \nu_{st}\left(\widetilde{\boldsymbol{\beta}}_{\psi}, \psi\right) - \nu_{st}\right) + \mathbf{0} \\
&= \operatorname{Cov}\left(\mathbf{L}^{\intercal}\mathbf{Y}, \nu_{st}\left(\widetilde{\boldsymbol{\beta}}_{\psi}, \psi\right) - \nu_{st}\right) = \mathbf{0},
\end{aligned}
$$

where $\mathbf{L}$ is an arbitrary nonrandom matrix such that $\operatorname{E}[\mathbf{L}^{\intercal}\mathbf{Y}] \equiv 0$. In particular,

$$
\operatorname{Cov}\left(\mathbf{Y} - \mathbf{X}^{\intercal}\widetilde{\boldsymbol{\beta}}(\psi), \phi_{st}\left(\widetilde{\boldsymbol{\beta}}_{\psi}, \psi\right) - \phi_{st}\right) = 0.
$$

Now we will show both $\phi_{st}\left(\widehat{\boldsymbol{\beta}}, \hat{\psi}\right)$ and $\phi_{st}\left(\widetilde{\boldsymbol{\beta}}_{\psi}, \psi\right)$ are location-equivariant. We call $d(\mathbf{Y})$ is location equivariant if $d(\mathbf{Y} + \mathbf{X}\boldsymbol{\alpha}) = d(\mathbf{Y}) + \mathbf{z}_{st}^{\intercal}\boldsymbol{\alpha}$ for all $\boldsymbol{\alpha}$ and $\mathbf{Y}$.

$$
\begin{aligned}
&\phi_{st}(\widetilde{\boldsymbol{\beta}}_{\psi}, \psi, \mathbf{Y} + \mathbf{X}\boldsymbol{\alpha}) \\
&= \mathbf{z}_{st}^{\intercal}\widetilde{\boldsymbol{\beta}}(\psi, \mathbf{Y} + \mathbf{X}\boldsymbol{\alpha}) + \psi\boldsymbol{\lambda}_{st}^{\intercal}\boldsymbol{\Sigma}^{-1}(\mathbf{Y} + \mathbf{X}\boldsymbol{\alpha} - \mathbf{X}\widetilde{\boldsymbol{\beta}}(\psi, \mathbf{Y} + \mathbf{X}\boldsymbol{\alpha})) \\
&= \mathbf{z}_{st}^{\intercal}(\mathbf{X}^{\intercal}\boldsymbol{\Sigma}(\psi)^{-1}\mathbf{X})^{-1}\mathbf{X}^{\intercal}\boldsymbol{\Sigma}(\psi)^{-1}(\mathbf{Y} + \mathbf{X}\boldsymbol{\alpha}) \\
&\quad + \psi\boldsymbol{\lambda}_{st}^{\intercal}\boldsymbol{\Sigma}(\psi)^{-1}(\mathbf{Y} + \mathbf{X}\boldsymbol{\alpha} - \mathbf{X}(\mathbf{X}^{\intercal}\boldsymbol{\Sigma}(\psi)^{-1}\mathbf{X})^{-1}\mathbf{X}^{\intercal}\boldsymbol{\Sigma}(\psi)^{-1}(\mathbf{Y} + \mathbf{X}\boldsymbol{\alpha})) \\
&= \mathbf{z}_{st}^{\intercal}(\mathbf{X}^{\intercal}\boldsymbol{\Sigma}(\psi)^{-1}\mathbf{X})^{-1}\mathbf{X}^{\intercal}\boldsymbol{\Sigma}(\psi)^{-1}\mathbf{Y} + \mathbf{z}_{st}^{\intercal}\boldsymbol{\alpha} \\
&\quad + \psi\boldsymbol{\lambda}_{st}^{\intercal}\boldsymbol{\Sigma}^{-1}(\mathbf{Y} - \mathbf{X}(\mathbf{X}^{\intercal}\boldsymbol{\Sigma}(\psi)^{-1}\mathbf{X})^{-1}\mathbf{X}^{\intercal}\boldsymbol{\Sigma}(\psi)^{-1}\mathbf{Y}) \\
&= \mathbf{z}_{st}^{\intercal}\widetilde{\boldsymbol{\beta}}(\psi, \mathbf{Y}) + \mathbf{z}_{st}^{\intercal}\boldsymbol{\alpha} + \psi\boldsymbol{\lambda}_{st}^{\intercal}\boldsymbol{\Sigma}(\psi)^{-1}(\mathbf{Y} - \mathbf{X}\widetilde{\boldsymbol{\beta}}(\psi, \mathbf{Y})) \\
&= \phi_{st}\left(\widetilde{\boldsymbol{\beta}}_{\psi}, \psi\right) + \mathbf{z}_{st}^{\intercal}\boldsymbol{\alpha}
\end{aligned}
$$

So $\phi_{st}\left(\widetilde{\boldsymbol{\beta}}_{\psi}, \psi\right)$ is location equivariant.

By similar arguments and the translation invariant property of $\hat{\psi}$, which means $\hat{\psi}(\mathbf{Y}+\mathbf{X}\boldsymbol{\alpha}) = \hat{\psi}(\mathbf{Y})$ for all $\boldsymbol{\alpha}$ and $\mathbf{Y}$, we can show $\phi_{st}\left(\widehat{\boldsymbol{\beta}}, \hat{\psi}\right)$ is location equivariant, too. So $\phi_{st}\left(\widehat{\boldsymbol{\beta}}, \hat{\psi}\right) - \phi_{st}\left(\widetilde{\boldsymbol{\beta}}_{\psi}, \psi\right)$ is translation invariant, and it can be reexpressed as a function of $\mathbf{Y} - \mathbf{X}^{\intercal}\widetilde{\boldsymbol{\beta}}(\psi)$. By normal distribution assumption, we have the independence of $\mathbf{Y} - \mathbf{X}^{\intercal}\widetilde{\boldsymbol{\beta}}(\psi)$ with $\phi_{st}\left(\widetilde{\boldsymbol{\beta}}_{\psi}, \psi\right) - \phi_{st}$, thus we have the independence of $\phi_{st}\left(\widehat{\boldsymbol{\beta}}, \hat{\psi}\right) - \phi_{st}\left(\widetilde{\boldsymbol{\beta}}_{\psi}, \psi\right)$ and $\phi_{st}\left(\widetilde{\boldsymbol{\beta}}_{\psi}, \psi\right) - \phi_{st}$. Thus,

$$\operatorname{Cov}\left(\phi_{st}\left(\widehat{\boldsymbol{\beta}}, \hat{\psi}\right) - \phi_{st}\left(\widetilde{\boldsymbol{\beta}}_{\psi}, \psi\right), \phi_{st}\left(\widetilde{\boldsymbol{\beta}}_{\psi}, \psi\right) - \phi_{st}\right) = 0. \tag{A.3.4}$$

(A.3.3) is easily achieved by subtracting (A.3.2) from (A.3.4). Next, we decompose the MSE of $\phi_{st}\left(\widehat{\boldsymbol{\beta}}, \hat{\psi}\right)$ into three parts and show them respectively.

$$\begin{aligned}
&\operatorname{MSE}\left(\phi_{st}\left(\widehat{\boldsymbol{\beta}}, \hat{\psi}\right)\right) \\
=& \mathrm{E}\left[\left(\phi_{st}\left(\widehat{\boldsymbol{\beta}}, \hat{\psi}\right) - \phi_{st}\right)^2\right] \\
=& \mathrm{E}\left[\left(\phi_{st}\left(\widehat{\boldsymbol{\beta}}, \hat{\psi}\right) - \phi_{st}\left(\widetilde{\boldsymbol{\beta}}_{\psi}, \psi\right) + \phi_{st}\left(\widetilde{\boldsymbol{\beta}}_{\psi}, \psi\right) - \phi_{st}\left(\boldsymbol{\beta}, \psi\right) + \phi_{st}\left(\boldsymbol{\beta}, \psi\right) - \phi_{st}\right)^2\right] \\
=& \mathrm{E}\left[\left(\phi_{st}\left(\boldsymbol{\beta}, \psi\right) - \phi_{st}\right)^2\right] + \mathrm{E}\left[\left(\phi_{st}\left(\widetilde{\boldsymbol{\beta}}_{\psi}, \psi\right) - \phi_{st}\left(\boldsymbol{\beta}, \psi\right)\right)^2\right] \\
&+ \mathrm{E}\left[\left(\phi_{st}\left(\widehat{\boldsymbol{\beta}}, \hat{\psi}\right) - \phi_{st}\left(\widetilde{\boldsymbol{\beta}}_{\psi}, \psi\right)\right)^2\right] \\
&+ 2\mathrm{E}\left[\left(\phi_{st}\left(\boldsymbol{\beta}, \psi\right) - \phi_{st}\right)\left(\phi_{st}\left(\widetilde{\boldsymbol{\beta}}_{\psi}, \psi\right) - \phi_{st}\left(\boldsymbol{\beta}, \psi\right)\right)\right] \\
&+ 2\mathrm{E}\left[\left(\phi_{st}\left(\boldsymbol{\beta}, \psi\right) - \phi_{st}\right)\left(\phi_{st}\left(\widehat{\boldsymbol{\beta}}, \hat{\psi}\right) - \phi_{st}\left(\widetilde{\boldsymbol{\beta}}_{\psi}, \psi\right)\right)\right] \\
&+ 2\mathrm{E}\left[\left(\phi_{st}\left(\widetilde{\boldsymbol{\beta}}_{\psi}, \psi\right) - \phi_{st}\left(\boldsymbol{\beta}, \psi\right)\right)\left(\phi_{st}\left(\widehat{\boldsymbol{\beta}}, \hat{\psi}\right) - \phi_{st}\left(\widetilde{\boldsymbol{\beta}}_{\psi}, \psi\right)\right)\right].
\end{aligned}$$

As $\mathrm{E}\left[\phi_{st}\left(\boldsymbol{\beta}, \psi\right) - \phi_{st}\right] = 0$ and $\mathrm{E}\left[\phi_{st}\left(\widetilde{\boldsymbol{\beta}}_{\psi}, \psi\right) - \phi_{st}\left(\boldsymbol{\beta}, \psi\right)\right] = 0$,

$$
\begin{aligned}
\mathrm{MSE}\left(\phi_{st}\left(\widehat{\boldsymbol{\beta}},\hat{\psi}\right)\right) &= \mathrm{E}\left[\left(\phi_{st}\left(\boldsymbol{\beta},\psi\right)-\phi_{st}\right)^{2}\right]+\mathrm{E}\left[\left(\phi_{st}\left(\widetilde{\boldsymbol{\beta}}_{\psi},\psi\right)-\phi_{st}\left(\boldsymbol{\beta},\psi\right)\right)^{2}\right] \\
&\quad+\mathrm{E}\left[\left(\phi_{st}\left(\widehat{\boldsymbol{\beta}},\hat{\psi}\right)-\phi_{st}\left(\widetilde{\boldsymbol{\beta}}_{\psi},\psi\right)\right)^{2}\right] \\
&\quad+2\mathrm{Cov}\left(\phi_{st}\left(\widetilde{\boldsymbol{\beta}}_{\psi},\psi\right)-\phi_{st}\left(\boldsymbol{\beta},\psi\right),\phi_{st}\left(\boldsymbol{\beta},\psi\right)-\phi_{st}\right) \\
&\quad+2\mathrm{Cov}\left(\phi_{st}\left(\widehat{\boldsymbol{\beta}},\hat{\psi}\right)-\phi_{st}\left(\widetilde{\boldsymbol{\beta}}_{\psi},\psi\right),\phi_{st}\left(\boldsymbol{\beta},\psi\right)-\phi_{st}\right) \\
&\quad+2\mathrm{Cov}\left(\phi_{st}\left(\widehat{\boldsymbol{\beta}},\hat{\psi}\right)-\phi_{st}\left(\widetilde{\boldsymbol{\beta}}_{\psi},\psi\right),\phi_{st}\left(\widetilde{\boldsymbol{\beta}}_{\psi},\psi\right)-\phi_{st}\left(\boldsymbol{\beta},\psi\right)\right) \\
&= \mathrm{E}\left[\left(\phi_{st}\left(\boldsymbol{\beta},\psi\right)-\phi_{st}\right)^{2}\right]+\mathrm{E}\left[\left(\phi_{st}\left(\widetilde{\boldsymbol{\beta}}_{\psi},\psi\right)-\phi_{st}\left(\boldsymbol{\beta},\psi\right)\right)^{2}\right] \\
&\quad+\mathrm{E}\left[\left(\phi_{st}\left(\widehat{\boldsymbol{\beta}},\hat{\psi}\right)-\phi_{st}\left(\widetilde{\boldsymbol{\beta}}_{\psi},\psi\right)\right)^{2}\right] \\
&= \dot{g}_{1st}+\dot{g}_{2st}+\dot{g}_{3st}.
\end{aligned}
$$

For each part,

$$
\begin{aligned}
\dot{g}_{1st} &= \mathrm{E}\left[\left(\phi_{st}\left(\boldsymbol{\beta},\psi\right)-\phi_{st}\right)^{2}\right] \\
&= \mathrm{E}\left[\left\{\left(\mathbf{z}_{st}^{\intercal}\boldsymbol{\beta}+\psi\boldsymbol{\lambda}_{st}^{\intercal}\boldsymbol{\Sigma}^{-1}(\psi)(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})\right)-\left(\mathbf{z}_{st}^{\intercal}\boldsymbol{\beta}+\nu_{st}\right)\right\}^{2}\right] \\
&= \mathrm{E}\left[\left(\psi\boldsymbol{\lambda}_{st}^{\intercal}\boldsymbol{\Sigma}^{-1}(\psi)(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})-\nu_{st}\right)^{2}\right] \\
&= \mathrm{E}\left[\left(\psi\boldsymbol{\lambda}_{st}^{\intercal}\boldsymbol{\Sigma}^{-1}(\psi)(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})-\boldsymbol{\lambda}_{st}^{\intercal}\boldsymbol{\nu}\right)\left(\psi\boldsymbol{\lambda}_{st}^{\intercal}\boldsymbol{\Sigma}^{-1}(\psi)(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})-\boldsymbol{\lambda}_{st}^{\intercal}\boldsymbol{\nu}\right)^{\intercal}\right] \\
&= \mathrm{E}\left[\psi\boldsymbol{\lambda}_{st}^{\intercal}\boldsymbol{\Sigma}^{-1}(\psi)(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})^{\intercal}\boldsymbol{\Sigma}^{-1}(\psi)\boldsymbol{\lambda}_{st}\psi\right] \\
&\quad-2\mathrm{E}\left[\psi\boldsymbol{\lambda}_{st}^{\intercal}\boldsymbol{\Sigma}^{-1}(\psi)(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})\boldsymbol{\nu}^{\intercal}\boldsymbol{\lambda}_{st}\right]+\mathrm{E}\left[\boldsymbol{\lambda}_{st}^{\intercal}\boldsymbol{\nu}\boldsymbol{\nu}^{\intercal}\boldsymbol{\lambda}_{st}\right] \\
&= \frac{\psi^{2}}{\psi+D_{st}}-2\frac{\psi^{2}}{\psi+D_{st}}+\psi \\
&= \frac{\psi D_{st}}{\psi+D_{st}},
\end{aligned}
$$

$$
\begin{aligned}
&\dot{g}_{2st} \\
&= \mathrm{E}\left[\left(\phi_{st}\left(\widetilde{\boldsymbol{\beta}}_{\psi}, \psi\right) - \phi_{st}\left(\boldsymbol{\beta}, \psi\right)\right)^2\right] \\
&= \mathrm{E}\left[\left\{\left(\mathbf{z}_{st}^{\intercal}\widetilde{\boldsymbol{\beta}} + \psi\boldsymbol{\lambda}_{st}^{\intercal}\boldsymbol{\Sigma}^{-1}(\psi)(\mathbf{Y} - \mathbf{X}\widetilde{\boldsymbol{\beta}})\right) - \left(\mathbf{z}_{st}^{\intercal}\boldsymbol{\beta} + \psi\boldsymbol{\lambda}_{st}^{\intercal}\boldsymbol{\Sigma}^{-1}(\psi)(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})\right)\right\}^2\right] \\
&= \mathrm{E}\left[\left\{\mathbf{z}_{st}^{\intercal}\left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) - \psi\boldsymbol{\lambda}_{st}^{\intercal}\boldsymbol{\Sigma}^{-1}(\psi)\mathbf{X}\left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\right\}^2\right] \\
&= \mathrm{E}\left[\left\{\left(\mathbf{z}_{st}^{\intercal} - \psi\boldsymbol{\lambda}_{st}^{\intercal}\boldsymbol{\Sigma}^{-1}(\psi)\mathbf{X}\right)\left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\right\}^2\right] \\
&= \mathrm{E}\left[\left(\boldsymbol{\lambda}_{st}^{\intercal}\mathbf{Z} - \psi\boldsymbol{\lambda}_{st}^{\intercal}\boldsymbol{\Sigma}^{-1}(\psi)\mathbf{X}\right)\left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^{\intercal}\left(\boldsymbol{\lambda}_{st}^{\intercal}\mathbf{Z} - \psi\boldsymbol{\lambda}_{st}^{\intercal}\boldsymbol{\Sigma}^{-1}(\psi)\mathbf{X}\right)^{\intercal}\right] \\
&= \left(\boldsymbol{\lambda}_{st}^{\intercal}\mathbf{Z} - \psi\boldsymbol{\lambda}_{st}^{\intercal}\boldsymbol{\Sigma}^{-1}(\psi)\mathbf{X}\right)\left(\mathbf{X}^{\intercal}\boldsymbol{\Sigma}^{-1}(\psi)\mathbf{X}\right)^{-1}\left(\boldsymbol{\lambda}_{st}^{\intercal}\mathbf{Z} - \psi\boldsymbol{\lambda}_{st}^{\intercal}\boldsymbol{\Sigma}^{-1}(\psi)\mathbf{X}\right)^{\intercal} \\
&= \left(\frac{\psi(\mathbf{z}_{st} - \mathbf{x}_{st})^{\intercal} + D_{st}\mathbf{z}_{st}^{\intercal}}{\psi + D_{st}}\right)\left[\sum_{u=1}^{m}(\psi + D_u)^{-1}\mathbf{x}_u\mathbf{x}_u^{\intercal}\right]^{-1}\left(\frac{\psi(\mathbf{z}_{st} - \mathbf{x}_{st})^{\intercal} + D_{st}\mathbf{z}_{st}^{\intercal}}{\psi + D_{st}}\right)^{\intercal},
\end{aligned}
$$

and

$$
\dot{g}_{3st} = \mathrm{E}\left[\left(\phi_{st}\left(\widehat{\boldsymbol{\beta}}, \hat{\psi}\right) - \phi_{st}\left(\widetilde{\boldsymbol{\beta}}_{\psi}, \psi\right)\right)^2\right].
$$

Following the proof of Theorem A.1 in Datta & Lahiri (2000), we can find that $\dot{g}_{3st} = g_{3st} + o(m^{-1})$, where g_{3st} can be found at Datta et al. (2005), and it has order $\mathcal{O}\left(m^{-1}\right)$. So

$$
\dot{g}_{3st} \simeq g_{3st} \simeq \frac{2D_{st}^2}{(\psi + D_{st})^3}\frac{1}{\sum_{u=1}^{m}(\psi + D_u)^2}
$$

□

A.4 Proof of MSE Approximation in Section 4.3.6

Proof. Recall that the moving average predictor is

$$
\widehat{\phi}_{st\mathrm{MA}} = \sum_{j=-K}^{K} a_j \phi_{s(t_1+j)t_2}\left(\widehat{\boldsymbol{\beta}}, \hat{\psi}\right),
$$

and the MSE is

$$
\begin{aligned}
\mathrm{MSE}\left(\widehat{\phi}_{st\mathrm{MA}}\right) &= \mathrm{E}\left[\left(\widehat{\phi}_{st\mathrm{MA}} - \phi_{st}\right)^2\right] \\
&= \mathrm{E}\left[\left(\sum_{j=-K}^{K} a_j \widehat{\phi}_{s(t_1+j)t_2} - \phi_{st_1t_2}\right)^2\right] \\
&= \mathrm{E}\left[\left(\sum_{j=-K}^{K} a_j \widehat{\phi}_{s(t_1+j)t_2} - \sum_{j=-K}^{K} a_j \phi_{st_1t_2}\right)^2\right] \\
&= \mathrm{E}\left[\sum_{j=-K}^{K} a_j^2 \left(\widehat{\phi}_{s(t_1+j)t_2} - \phi_{st_1t_2}\right)^2\right] \\
&\quad + \mathrm{E}\left[\sum\sum_{i\neq j} a_i a_j \left(\widehat{\phi}_{s(t_1+i)t_2} - \phi_{st_1t_2}\right)\left(\widehat{\phi}_{s(t_1+j)t_2} - \phi_{st_1t_2}\right)\right] \\
&= \sum_{j=-K}^{K} a_j^2 \mathrm{E}\left[\left(\widehat{\phi}_{s(t_1+j)t_2} - \phi_{st_1t_2}\right)^2\right] \\
&\quad + \sum\sum_{i\neq j} a_i a_j \mathrm{E}\left[\left(\widehat{\phi}_{s(t_1+i)t_2} - \phi_{st_1t_2}\right)\left(\widehat{\phi}_{s(t_1+j)t_2} - \phi_{st_1t_2}\right)\right]. \qquad \text{(A.4.1)}
\end{aligned}
$$

We check the two expectations separately. Firstly,

$$
\begin{aligned}
\mathrm{E}\left[\left(\widehat{\phi}_{s(t_1+j)t_2} - \phi_{st_1t_2}\right)^2\right] &= \mathrm{E}\left[\left(\widehat{\phi}_{s(t_1+j)t_2} - \phi_{s(t_1+j)t_2} + \phi_{s(t_1+j)t_2} - \phi_{st_1t_2}\right)^2\right] \\
&= \mathrm{E}\left[\left(\widehat{\phi}_{s(t_1+j)t_2} - \phi_{s(t_1+j)t_2}\right)^2\right] + \mathrm{E}\left[\left(\phi_{s(t_1+j)t_2} - \phi_{st_1t_2}\right)^2\right] \\
&\quad + 2\mathrm{E}\left[\left(\widehat{\phi}_{s(t_1+j)t_2} - \phi_{s(t_1+j)t_2}\right)\left(\phi_{s(t_1+j)t_2} - \phi_{st_1t_2}\right)\right],
\end{aligned}
$$

where the expression of $\mathrm{E}\left[\left(\widehat{\phi}_{s(t_1+j)t_2} - \phi_{s(t_1+j)t_2}\right)^2\right]$ can be found at Section 4.3.3. To simplify the notation, we will use t and t' to denote $(t_1+j)t_2$ and t_1t_2 respectively in the cross product term above.

$$\begin{aligned}
&\mathrm{E}\left[\left(\widehat{\phi}_{s(t_1+j)t_2}-\phi_{s(t_1+j)t_2}\right)\left(\phi_{s(t_1+j)t_2}-\phi_{st_1t_2}\right)\right]\\
&=\mathrm{E}\left[\left(\phi_{st}\left(\widehat{\boldsymbol{\beta}},\hat{\psi}\right)-\phi_{st}\right)\left(\phi_{st}-\phi_{st'}\right)\right]\\
&=\mathrm{E}\left[\left(\phi_{st}\left(\widehat{\boldsymbol{\beta}},\hat{\psi}\right)-\phi_{st}\left(\widetilde{\boldsymbol{\beta}}_{\psi},\psi\right)+\phi_{st}\left(\widetilde{\boldsymbol{\beta}}_{\psi},\psi\right)-\phi_{st}\left(\boldsymbol{\beta},\psi\right)+\phi_{st}\left(\boldsymbol{\beta},\psi\right)-\phi_{st}\right)\left(\phi_{st}-\phi_{st'}\right)\right]\\
&=\mathrm{E}\left[\left(\phi_{st}\left(\boldsymbol{\beta},\psi\right)-\phi_{st}\right)\left(\phi_{st}-\phi_{st'}\right)\right]+\mathrm{E}\left[\left(\phi_{st}\left(\widetilde{\boldsymbol{\beta}}_{\psi},\psi\right)-\phi_{st}\left(\boldsymbol{\beta},\psi\right)\right)\left(\phi_{st}-\phi_{st'}\right)\right]\\
&\quad+\mathrm{E}\left[\left(\phi_{st}\left(\widehat{\boldsymbol{\beta}},\hat{\psi}\right)-\phi_{st}\left(\widetilde{\boldsymbol{\beta}}_{\psi},\psi\right)\right)\left(\phi_{st}-\phi_{st'}\right)\right]\\
&=\mathrm{E}\left[\left(\phi_{st}\left(\widehat{\boldsymbol{\beta}},\hat{\psi}\right)-\phi_{st}\left(\widetilde{\boldsymbol{\beta}}_{\psi},\psi\right)\right)\left(\phi_{st}-\phi_{st'}\right)\right]\\
&=\mathrm{E}\left[\left(\phi_{st}-\phi_{st'}\right)\mathrm{E}\left[\left(\phi_{st}\left(\widehat{\boldsymbol{\beta}},\hat{\psi}\right)-\phi_{st}\left(\widetilde{\boldsymbol{\beta}}_{\psi},\psi\right)\right)\middle|\,\phi_{st},\phi_{st'}\right]\right].
\end{aligned}$$

By Cauchy–Schwarz inequality,

$$\begin{aligned}
&\mathrm{E}\left[\left(\phi_{st}\left(\widehat{\boldsymbol{\beta}},\hat{\psi}\right)-\phi_{st}\left(\widetilde{\boldsymbol{\beta}}_{\psi},\psi\right)\right)\middle|\,\phi_{st},\phi_{st'}\right]\\
&\leq\left\{\mathrm{E}\left[\left(\phi_{st}\left(\widehat{\boldsymbol{\beta}},\hat{\psi}\right)-\phi_{st}\left(\widetilde{\boldsymbol{\beta}}_{\psi},\psi\right)\right)^2\middle|\,\phi_{st},\phi_{st'}\right]\right\}^{1/2}=\mathcal{O}\left(m^{-1/2}\right),
\end{aligned}$$

which comes from the order of $\dot{g}_{3st}$ in Section A.3. So

$$\mathrm{E}\left[\left(\widehat{\phi}_{s(t_1+j)t_2}-\phi_{s(t_1+j)t_2}\right)\left(\phi_{s(t_1+j)t_2}-\phi_{st_1t_2}\right)\right]=\mathcal{O}\left(m^{-1/2}\right), \tag{A.4.2}$$

as $\mathrm{E}\left[\phi_{st}-\phi_{st'}\right]$ is $\mathcal{O}\left(1\right)$. Thus we have

$$\mathrm{E}\left[\left(\widehat{\phi}_{s(t_1+j)t_2}-\phi_{st_1t_2}\right)^2\right]=\mathrm{MSE}\left(\widehat{\phi}_{s(t_1+j)t_2}\right)+\mathrm{E}\left[\left(\phi_{s(t_1+j)t_2}-\phi_{st_1t_2}\right)^2\right]+\mathcal{O}\left(m^{-1/2}\right), \tag{A.4.3}$$

Next, we look at the cross product term in (A.4.1).

$$
\begin{aligned}
&\mathrm{E}\left[\left(\widehat{\phi}_{s(t_1+i)t_2}-\phi_{st_1t_2}\right)\left(\widehat{\phi}_{s(t_1+j)t_2}-\phi_{st_1t_2}\right)\right]\\
=&\,\mathrm{E}\left[\left(\widehat{\phi}_{s(t_1+i)t_2}-\phi_{s(t_1+i)t_2}+\phi_{s(t_1+i)t_2}-\phi_{st_1t_2}\right)\left(\widehat{\phi}_{s(t_1+j)t_2}-\phi_{s(t_1+j)t_2}+\phi_{s(t_1+j)t_2}-\phi_{st_1t_2}\right)\right]\\
=&\,\mathrm{E}\left[\left(\widehat{\phi}_{s(t_1+i)t_2}-\phi_{s(t_1+i)t_2}\right)\left(\widehat{\phi}_{s(t_1+j)t_2}-\phi_{s(t_1+j)t_2}\right)\right]\\
&+\mathrm{E}\left[\left(\widehat{\phi}_{s(t_1+i)t_2}-\phi_{s(t_1+i)t_2}\right)\left(\phi_{s(t_1+j)t_2}-\phi_{st_1t_2}\right)\right]\\
&+\mathrm{E}\left[\left(\widehat{\phi}_{s(t_1+j)t_2}-\phi_{s(t_1+j)t_2}\right)\left(\phi_{s(t_1+i)t_2}-\phi_{st_1t_2}\right)\right]\\
&+\mathrm{E}\left[\left(\phi_{s(t_1+i)t_2}-\phi_{st_1t_2}\right)\left(\phi_{s(t_1+j)t_2}-\phi_{st_1t_2}\right)\right]\\
=&\,A+B+C+D.
\end{aligned}
$$

B and C have same order as (A.4.2). To simplify the notation, we use t and t' to denote $(t_1+i)t_2$ and $(t_1+j)t_2$ respectively. We only need check A now.

$$
\begin{aligned}
&\mathrm{E}\left[\left(\widehat{\phi}_{s(t_1+i)t_2}-\phi_{s(t_1+i)t_2}\right)\left(\widehat{\phi}_{s(t_1+j)t_2}-\phi_{s(t_1+j)t_2}\right)\right]\\
=&\,\mathrm{E}\left[\left(\phi_{st}\left(\widehat{\boldsymbol{\beta}},\hat{\psi}\right)-\phi_{st}\right)\left(\phi_{st'}\left(\widehat{\boldsymbol{\beta}},\hat{\psi}\right)-\phi_{st'}\right)\right]\\
=&\,\mathrm{E}\Bigg[\left(\phi_{st}\left(\widehat{\boldsymbol{\beta}},\hat{\psi}\right)-\phi_{st}\left(\widetilde{\boldsymbol{\beta}}_{\psi},\psi\right)+\phi_{st}\left(\widetilde{\boldsymbol{\beta}}_{\psi},\psi\right)-\phi_{st}\left(\boldsymbol{\beta},\psi\right)+\phi_{st}\left(\boldsymbol{\beta},\psi\right)-\phi_{st}\right)\\
&\times\left(\phi_{st'}\left(\widehat{\boldsymbol{\beta}},\hat{\psi}\right)-\phi_{st'}\left(\widetilde{\boldsymbol{\beta}}_{\psi},\psi\right)+\phi_{st'}\left(\widetilde{\boldsymbol{\beta}}_{\psi},\psi\right)-\phi_{st'}\left(\boldsymbol{\beta},\psi\right)+\phi_{st'}\left(\boldsymbol{\beta},\psi\right)-\phi_{st'}\right)\Bigg]\\
=&\,\mathrm{E}\left[\left(\phi_{st}\left(\boldsymbol{\beta},\psi\right)-\phi_{st}\right)\left(\phi_{st'}\left(\boldsymbol{\beta},\psi\right)-\phi_{st'}\right)\right]\\
&+\mathrm{E}\left[\left(\phi_{st}\left(\widetilde{\boldsymbol{\beta}}_{\psi},\psi\right)-\phi_{st}\left(\boldsymbol{\beta},\psi\right)\right)\left(\phi_{st'}\left(\widetilde{\boldsymbol{\beta}}_{\psi},\psi\right)-\phi_{st'}\left(\boldsymbol{\beta},\psi\right)\right)\right]\\
&+\mathrm{E}\left[\left(\phi_{st}\left(\widehat{\boldsymbol{\beta}},\hat{\psi}\right)-\phi_{st}\left(\widetilde{\boldsymbol{\beta}}_{\psi},\psi\right)\right)\left(\phi_{st'}\left(\widehat{\boldsymbol{\beta}},\hat{\psi}\right)-\phi_{st'}\left(\widetilde{\boldsymbol{\beta}}_{\psi},\psi\right)\right)\right]\\
&+\mathrm{E}\left[\left(\phi_{st}\left(\boldsymbol{\beta},\psi\right)-\phi_{st}\right)\left(\phi_{st'}\left(\widetilde{\boldsymbol{\beta}}_{\psi},\psi\right)-\phi_{st'}\left(\boldsymbol{\beta},\psi\right)\right)\right]\\
&+\mathrm{E}\left[\left(\phi_{st}\left(\boldsymbol{\beta},\psi\right)-\phi_{st}\right)\left(\phi_{st'}\left(\widehat{\boldsymbol{\beta}},\hat{\psi}\right)-\phi_{st'}\left(\widetilde{\boldsymbol{\beta}}_{\psi},\psi\right)\right)\right]\\
&+\mathrm{E}\left[\left(\phi_{st}\left(\widetilde{\boldsymbol{\beta}}_{\psi},\psi\right)-\phi_{st}\left(\boldsymbol{\beta},\psi\right)\right)\left(\phi_{st'}\left(\widehat{\boldsymbol{\beta}},\hat{\psi}\right)-\phi_{st'}\left(\widetilde{\boldsymbol{\beta}}_{\psi},\psi\right)\right)\right]\\
&+\mathrm{E}\left[\left(\phi_{st'}\left(\boldsymbol{\beta},\psi\right)-\phi_{st'}\right)\left(\phi_{st}\left(\widetilde{\boldsymbol{\beta}}_{\psi},\psi\right)-\phi_{st}\left(\boldsymbol{\beta},\psi\right)\right)\right]\\
&+\mathrm{E}\left[\left(\phi_{st'}\left(\boldsymbol{\beta},\psi\right)-\phi_{st'}\right)\left(\phi_{st}\left(\widehat{\boldsymbol{\beta}},\hat{\psi}\right)-\phi_{st}\left(\widetilde{\boldsymbol{\beta}}_{\psi},\psi\right)\right)\right]\\
&+\mathrm{E}\left[\left(\phi_{st'}\left(\widetilde{\boldsymbol{\beta}}_{\psi},\psi\right)-\phi_{st'}\left(\boldsymbol{\beta},\psi\right)\right)\left(\phi_{st}\left(\widehat{\boldsymbol{\beta}},\hat{\psi}\right)-\phi_{st}\left(\widetilde{\boldsymbol{\beta}}_{\psi},\psi\right)\right)\right].
\end{aligned}
$$

As $\mathrm{E}\left[\phi_{st}\left(\boldsymbol{\beta},\psi\right)-\phi_{st}\right]=0$, $\mathrm{E}\left[\phi_{st}\left(\widetilde{\boldsymbol{\beta}}_{\psi},\psi\right)-\phi_{st}\left(\boldsymbol{\beta},\psi\right)\right]=0$, $\mathrm{E}\left[\phi_{st'}\left(\boldsymbol{\beta},\psi\right)-\phi_{st'}\right]=0$ and $\mathrm{E}\left[\phi_{st'}\left(\widetilde{\boldsymbol{\beta}}_{\psi},\psi\right)-\phi_{st'}\left(\boldsymbol{\beta},\psi\right)\right]=0$,

$$\begin{aligned}
&\mathrm{E}\left[\left(\widehat{\phi}_{s(t_1+i)t_2}-\phi_{s(t_1+i)t_2}\right)\left(\widehat{\phi}_{s(t_1+j)t_2}-\phi_{s(t_1+j)t_2}\right)\right]\\
=&\,\mathrm{E}\left[\left(\phi_{st}\left(\boldsymbol{\beta},\psi\right)-\phi_{st}\right)\left(\phi_{st'}\left(\boldsymbol{\beta},\psi\right)-\phi_{st'}\right)\right]\\
&+\mathrm{E}\left[\left(\phi_{st}\left(\widetilde{\boldsymbol{\beta}}_{\psi},\psi\right)-\phi_{st}\left(\boldsymbol{\beta},\psi\right)\right)\left(\phi_{st'}\left(\widetilde{\boldsymbol{\beta}}_{\psi},\psi\right)-\phi_{st'}\left(\boldsymbol{\beta},\psi\right)\right)\right]\\
&+\mathrm{E}\left[\left(\phi_{st}\left(\widehat{\boldsymbol{\beta}},\hat{\psi}\right)-\phi_{st}\left(\widetilde{\boldsymbol{\beta}}_{\psi},\psi\right)\right)\left(\phi_{st'}\left(\widehat{\boldsymbol{\beta}},\hat{\psi}\right)-\phi_{st'}\left(\widetilde{\boldsymbol{\beta}}_{\psi},\psi\right)\right)\right]\\
&+\mathrm{Cov}\left(\phi_{st}\left(\widetilde{\boldsymbol{\beta}}_{\psi},\psi\right)-\phi_{st}\left(\boldsymbol{\beta},\psi\right),\phi_{st'}\left(\boldsymbol{\beta},\psi\right)-\phi_{st'}\right)\\
&+\mathrm{Cov}\left(\phi_{st}\left(\widehat{\boldsymbol{\beta}},\hat{\psi}\right)-\phi_{st}\left(\widetilde{\boldsymbol{\beta}}_{\psi},\psi\right),\phi_{st'}\left(\boldsymbol{\beta},\psi\right)-\phi_{st'}\right)\\
&+\mathrm{Cov}\left(\phi_{st}\left(\widehat{\boldsymbol{\beta}},\hat{\psi}\right)-\phi_{st}\left(\widetilde{\boldsymbol{\beta}}_{\psi},\psi\right),\phi_{st'}\left(\widetilde{\boldsymbol{\beta}}_{\psi},\psi\right)-\phi_{st'}\left(\boldsymbol{\beta},\psi\right)\right)\\
&+\mathrm{Cov}\left(\phi_{st'}\left(\widetilde{\boldsymbol{\beta}}_{\psi},\psi\right)-\phi_{st'}\left(\boldsymbol{\beta},\psi\right),\phi_{st}\left(\boldsymbol{\beta},\psi\right)-\phi_{st}\right)\\
&+\mathrm{Cov}\left(\phi_{st'}\left(\widehat{\boldsymbol{\beta}},\hat{\psi}\right)-\phi_{st'}\left(\widetilde{\boldsymbol{\beta}}_{\psi},\psi\right),\phi_{st}\left(\boldsymbol{\beta},\psi\right)-\phi_{st}\right)\\
&+\mathrm{Cov}\left(\phi_{st'}\left(\widehat{\boldsymbol{\beta}},\hat{\psi}\right)-\phi_{st'}\left(\widetilde{\boldsymbol{\beta}}_{\psi},\psi\right),\phi_{st}\left(\widetilde{\boldsymbol{\beta}}_{\psi},\psi\right)-\phi_{st}\left(\boldsymbol{\beta},\psi\right)\right).
\end{aligned}$$

By arguments in (A.3), the six covariances are zero. So

$$\begin{aligned}
&\mathrm{E}\left[\left(\phi_{st}\left(\widehat{\boldsymbol{\beta}},\hat{\psi}\right)-\phi_{st}\right)\left(\phi_{st'}\left(\widehat{\boldsymbol{\beta}},\hat{\psi}\right)-\phi_{st'}\right)\right]\\
=&\,\mathrm{E}\left[\left(\phi_{st}\left(\boldsymbol{\beta},\psi\right)-\phi_{st}\right)\left(\phi_{st'}\left(\boldsymbol{\beta},\psi\right)-\phi_{st'}\right)\right]\\
&+\mathrm{E}\left[\left(\phi_{st}\left(\widetilde{\boldsymbol{\beta}}_{\psi},\psi\right)-\phi_{st}\left(\boldsymbol{\beta},\psi\right)\right)\left(\phi_{st'}\left(\widetilde{\boldsymbol{\beta}}_{\psi},\psi\right)-\phi_{st'}\left(\boldsymbol{\beta},\psi\right)\right)\right]\\
&+\mathrm{E}\left[\left(\phi_{st}\left(\widehat{\boldsymbol{\beta}},\hat{\psi}\right)-\phi_{st}\left(\widetilde{\boldsymbol{\beta}}_{\psi},\psi\right)\right)\left(\phi_{st'}\left(\widehat{\boldsymbol{\beta}},\hat{\psi}\right)-\phi_{st'}\left(\widetilde{\boldsymbol{\beta}}_{\psi},\psi\right)\right)\right]\\
=&\,\dot{g}_{1stt'}+\dot{g}_{2stt'}+\dot{g}_{3stt'}.
\end{aligned}$$

For each part,

$$
\begin{aligned}
\dot{g}_{1stt'} &= \mathrm{E}\left[\left(\phi_{st}\left(\widehat{\boldsymbol{\beta}},\hat{\psi}\right)-\phi_{st}\right)\left(\phi_{st'}\left(\widehat{\boldsymbol{\beta}},\hat{\psi}\right)-\phi_{st'}\right)\right] \\
&= \mathrm{E}\left[\left\{\left(\mathbf{z}_{st}^{\intercal}\boldsymbol{\beta}+\psi\boldsymbol{\lambda}_{st}^{\intercal}\boldsymbol{\Sigma}^{-1}(\psi)(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})\right)-(\mathbf{z}_{st}^{\intercal}\boldsymbol{\beta}+\nu_{st})\right\}\right. \\
&\quad\left.\times\left\{\left(\mathbf{z}_{st'}^{\intercal}\boldsymbol{\beta}+\psi\boldsymbol{\lambda}_{st'}^{\intercal}\boldsymbol{\Sigma}^{-1}(\psi)(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})\right)-(\mathbf{z}_{st'}^{\intercal}\boldsymbol{\beta}+\nu_{st'})\right\}\right] \\
&= \mathrm{E}\left[\left(\psi\boldsymbol{\lambda}_{st}^{\intercal}\boldsymbol{\Sigma}^{-1}(\psi)(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})-\nu_{st}\right)\left(\psi\boldsymbol{\lambda}_{st'}^{\intercal}\boldsymbol{\Sigma}^{-1}(\psi)(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})-\nu_{st'}\right)\right] \\
&= \mathrm{E}\left[\left(\psi\boldsymbol{\lambda}_{st}^{\intercal}\boldsymbol{\Sigma}^{-1}(\psi)(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})-\boldsymbol{\lambda}_{st}^{\intercal}\boldsymbol{\nu}\right)\left(\psi\boldsymbol{\lambda}_{st'}^{\intercal}\boldsymbol{\Sigma}^{-1}(\psi)(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})-\boldsymbol{\lambda}_{st'}^{\intercal}\boldsymbol{\nu}\right)^{\intercal}\right] \\
&= \mathrm{E}\left[\psi\boldsymbol{\lambda}_{st}^{\intercal}\boldsymbol{\Sigma}^{-1}(\psi)(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})^{\intercal}\boldsymbol{\Sigma}^{-1}(\psi)\boldsymbol{\lambda}_{st'}\psi\right] \\
&\quad-\mathrm{E}\left[\psi\boldsymbol{\lambda}_{st}^{\intercal}\boldsymbol{\Sigma}^{-1}(\psi)(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})\boldsymbol{\nu}^{\intercal}\boldsymbol{\lambda}_{st'}\right] \\
&\quad-\mathrm{E}\left[\boldsymbol{\lambda}_{st}^{\intercal}\boldsymbol{\nu}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})^{\intercal}\boldsymbol{\Sigma}^{-1}(\psi)\boldsymbol{\lambda}_{st'}\psi\right]+\mathrm{E}\left[\boldsymbol{\lambda}_{st}^{\intercal}\boldsymbol{\nu}\boldsymbol{\nu}^{\intercal}\boldsymbol{\lambda}_{st'}\right] \\
&= 0.
\end{aligned}
$$

The last equality holds as t and t' are different time and the covariance matrix is diagonal by independence of random effects and sampling error.

$$
\begin{aligned}
&\dot{g}_{2stt'} \\
&= \mathrm{E}\left[\left(\phi_{st}\left(\widetilde{\boldsymbol{\beta}}_{\psi},\psi\right)-\phi_{st}\left(\boldsymbol{\beta},\psi\right)\right)\left(\phi_{st'}\left(\widetilde{\boldsymbol{\beta}}_{\psi},\psi\right)-\phi_{st'}\left(\boldsymbol{\beta},\psi\right)\right)\right] \\
&= \mathrm{E}\left[\left\{\left(\mathbf{z}_{st}^{\intercal}\widetilde{\boldsymbol{\beta}}+\psi\boldsymbol{\lambda}_{st}^{\intercal}\boldsymbol{\Sigma}^{-1}(\psi)(\mathbf{Y}-\mathbf{X}\widetilde{\boldsymbol{\beta}})\right)-\left(\mathbf{z}_{st}^{\intercal}\boldsymbol{\beta}+\psi\boldsymbol{\lambda}_{st}^{\intercal}\boldsymbol{\Sigma}^{-1}(\psi)(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})\right)\right\}\right. \\
&\quad\left.\times\left\{\left(\mathbf{z}_{st'}^{\intercal}\widetilde{\boldsymbol{\beta}}+\psi\boldsymbol{\lambda}_{st'}^{\intercal}\boldsymbol{\Sigma}^{-1}(\psi)(\mathbf{Y}-\mathbf{X}\widetilde{\boldsymbol{\beta}})\right)-\left(\mathbf{z}_{st'}^{\intercal}\boldsymbol{\beta}+\psi\boldsymbol{\lambda}_{st'}^{\intercal}\boldsymbol{\Sigma}^{-1}(\psi)(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})\right)\right\}\right] \\
&= \mathrm{E}\left[\left\{\mathbf{z}_{st}^{\intercal}\left(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)-\psi\boldsymbol{\lambda}_{st}^{\intercal}\boldsymbol{\Sigma}^{-1}(\psi)\mathbf{X}\left(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\right\}\right. \\
&\quad\left.\times\left\{\mathbf{z}_{st'}^{\intercal}\left(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)-\psi\boldsymbol{\lambda}_{st'}^{\intercal}\boldsymbol{\Sigma}^{-1}(\psi)\mathbf{X}\left(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\right\}\right] \\
&= \mathrm{E}\left[\left\{\left(\mathbf{z}_{st}^{\intercal}-\psi\boldsymbol{\lambda}_{st}^{\intercal}\boldsymbol{\Sigma}^{-1}(\psi)\mathbf{X}\right)\left(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\right\}\left\{\left(\mathbf{z}_{st'}^{\intercal}-\psi\boldsymbol{\lambda}_{st'}^{\intercal}\boldsymbol{\Sigma}^{-1}(\psi)\mathbf{X}\right)\left(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\right\}\right] \\
&= \mathrm{E}\left[\left(\boldsymbol{\lambda}_{st}^{\intercal}\mathbf{Z}-\psi\boldsymbol{\lambda}_{st}^{\intercal}\boldsymbol{\Sigma}^{-1}(\psi)\mathbf{X}\right)\left(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\left(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)^{\intercal}\left(\boldsymbol{\lambda}_{st'}^{\intercal}\mathbf{Z}-\psi\boldsymbol{\lambda}_{st'}^{\intercal}\boldsymbol{\Sigma}^{-1}(\psi)\mathbf{X}\right)^{\intercal}\right] \\
&= \left(\boldsymbol{\lambda}_{st}^{\intercal}\mathbf{Z}-\psi\boldsymbol{\lambda}_{st}^{\intercal}\boldsymbol{\Sigma}^{-1}(\psi)\mathbf{X}\right)\left(\mathbf{X}^{\intercal}\boldsymbol{\Sigma}^{-1}(\psi)\mathbf{X}\right)^{-1}\left(\boldsymbol{\lambda}_{st'}^{\intercal}\mathbf{Z}-\psi\boldsymbol{\lambda}_{st'}^{\intercal}\boldsymbol{\Sigma}^{-1}(\psi)\mathbf{X}\right)^{\intercal} \\
&= \left(\frac{\psi(\mathbf{z}_{st}-\mathbf{x}_{st})^{\intercal}+D_{st}\mathbf{z}_{st}^{\intercal}}{\psi+D_{st}}\right)\left[\sum_{u=1}^{m}(\psi+D_u)^{-1}\mathbf{x}_u\mathbf{x}_u^{\intercal}\right]^{-1}\left(\frac{\psi(\mathbf{z}_{st'}-\mathbf{x}_{st'})^{\intercal}+D_{st'}\mathbf{z}_{st'}^{\intercal}}{\psi+D_{st'}}\right)^{\intercal},
\end{aligned}
$$

and

$$
\begin{aligned}
\dot{g}_{3stt'} &= \mathrm{E}\left[\left(\phi_{st}\left(\widehat{\boldsymbol{\beta}},\hat{\psi}\right)-\phi_{st}\left(\widetilde{\boldsymbol{\beta}}_{\psi},\psi\right)\right)\left(\phi_{st'}\left(\widehat{\boldsymbol{\beta}},\hat{\psi}\right)-\phi_{st'}\left(\widetilde{\boldsymbol{\beta}}_{\psi},\psi\right)\right)\right] \\
&\simeq \frac{2D_{st}^2}{(\psi+D_{st})^3}\frac{1}{\sum_{u=1}^{m}(\psi+D_u)^2}.
\end{aligned}
$$

So

$$
\begin{aligned}
&\mathrm{E}\left[\left(\widehat{\phi}_{s(t_1+i)t_2}-\phi_{st_1t_2}\right)\left(\widehat{\phi}_{s(t_1+j)t_2}-\phi_{st_1t_2}\right)\right] \\
&= \mathrm{E}\left[\left(\phi_{s(t_1+i)t_2}-\phi_{st_1t_2}\right)\left(\phi_{s(t_1+j)t_2}-\phi_{st_1t_2}\right)\right]+\mathcal{O}\left(m^{-1/2}\right). \qquad \text{(A.4.4)}
\end{aligned}
$$

Then by combining (A.4.3) and (A.4.4), we have

$$
\begin{aligned}
\mathrm{MSE}\left(\widehat{\phi}_{st\mathrm{MA}}\right) &= \sum_{j=-K}^{K} a_j^2\left\{\mathrm{MSE}\left(\widehat{\phi}_{s(t_1+j)t_2}\right)+\mathrm{E}\left[\left(\phi_{s(t_1+j)t_2}-\phi_{st_1t_2}\right)^2\right]\right\} \\
&\quad+\sum\sum_{i\neq j} a_i a_j \mathrm{E}\left[\left(\phi_{s(t_1+i)t_2}-\phi_{st_1t_2}\right)\left(\phi_{s(t_1+j)t_2}-\phi_{st_1t_2}\right)\right]+\mathcal{O}\left(m^{-1/2}\right).
\end{aligned}
\qquad \text{(A.4.5)}
$$

□