POST-SELECTION INFERENCE

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Alexandra Bolotskikh

February 2016

POST-SELECTION INFERENCE

Alexandra Bolotskikh, Ph.D.

Cornell University 2016

Researchers are often interested in making inference on one or a few "best" treatment out of p given treatments. This problem is referred to as post-selection inference. Substantial research has been done on constructing point estimates and confidence sets for a multivariate normal mean vector, but there are very few works on inference after selection. Often, out of all available estimates, the ones that are minimax and/or admissible are preferred. It was shown by Sackrowitz & Samuel-Cahn (1986) that X_1, the first order statistic, is minimax for estimating the selected mean for $p \leq 3$, but it is not minimax for $p > 3$, but the question whether it is admissible was still open.

In Chapter 1, following the arguments of Berger (1976a) and Maruyama (2009) we prove that X_1 is admissible for $p < 4$ and some bias correction is needed for $p \geq 4$. The bias corrected estimate, a generalized Bayes estimate, will be admissible for $p \geq 4$. We also provide a comparison of this admissible estimator under the horseshoe prior and other estimators proposed in the literature, including the estimator proposed in Reid & Tibshirani (2014) and a generalized Bayes estimator under the horseshoe prior.

Naive confidence sets under selection fail to maintain the nominal coverage probability. In Chapter 2, by approximating the coverage probability of the naive set, we derive two confidence sets that maintain coverage probability. The first set is straightforward to construct, but it results in a large conservative set with coverage probabilities close to one. The second set is more involved to construct, but it is significantly improve over the first one.

In Chapter 3 we explore minimaxity and admissibility of the naive set under selection. If we require a set to maintain coverage probability, we show that it is only minimax for $p = 2$. Almost admissibility is shown for $p = 2$. The rest of the chapter is dedicated to investigating potential dominating sets.

BIOGRAPHICAL SKETCH

Alexandra Bolotskikh was born in Ukraine in 1988. After graduation from high school she enrolled at the Saint-Petersburg State University, where she received a Bachelor's degree in Applied Mathematics and Physics in 2009. Her Bachelor's degree work was on spectral analysis of Dow Jones index, which was her first exposure to statistics.

In August 2009 Alexandra moved to U.S. for her graduate degree. She entered graduated program in the Department of Statistics at the University of Florida, where she received her Masters degree in 2012. After transferring to Cornell, she earned her Ph.D. in Statistics in 2015 under the supervision of Dr. Martin Wells.

To my family.

ACKNOWLEDGEMENTS

I am very grateful to my advisor, Dr. Martin Wells, for his thoughtful guidance and support, and for his enthusiasm for the problems we worked on. Without his consistent encouragement, it would probably not be possible to successfully finish my studies and graduate from Cornell.

I would like to express my gratitude to my first advisor at the University of Florida, Dr. George Casella, who believed in me and introduced me to the topic of post-selection inference which became the subject of my dissertation.

I would also like to thank my graduate committee members: James Booth and Jacob Bien for their guidance.

To my friends from University of Florida: Claudio Fuentes, Daniel Taylor and Tavis Abrahamsen; and to my friends from Cornell: James Li, Didier Chetelat, David Sinclair, James Davis and Mathav Murugan, who were always there for me, thank you!

Lastly, and most importantly, I would like to thank my parents, Galina and Vladimir Bolotskikh, for supporting my decision to come to graduate school in the U.S. and for their unconditional love and support, and my sister Yulia who supported me every step of the way and taught me to never give up.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Researchers are often interested in making inference on one or a few "best" treatments out of p given treatments. Some examples include: a clinical trial where there are multiple drugs tested, a medical researcher might want to choose the drug or drugs with the best performance; a farmer who used several fertilizers last year, now wants to choose the ones that produced maximum yield; and an animal breeder has hundreds of cows, and wants to choose the ones with the highest milk production. More recently selection problems for choosing a few genes with the highest expression values out of thousands available genes have important scientific consequences. Each of these applications is an example of inference after selection. Our objective is to choose the treatment with the largest sample mean and after selection is made, construct point estimates and confidence sets for the mean of the selected population.

For point estimation, if we use the naive estimator as if there were no prior selection made, we get a biased estimator. This bias can be especially large if the dimension of the parameter space is large or means are close together. The problem of improving upon the usual intuitive estimators is difficult, but extensive research has been done on this topic. For example, under normality Guttman & Tiao (1964) used a Bayesian approach to choose the best treatment by maximizing expected utility function; Hwang (1993) provided an empirical Bayes estimator and showed that it was better than the usual estimator in terms of the Bayes risk with respect to any normal prior. Hwang (1993) also provided a comparison with an estimator, T_2, of Cohen & Sackrowitz (1982). Hwang's estimator performs favorably, but comparison might not be fair since T_2 was not designed specifically for iid priors, which were assumed by Hwang (1993).

In this work we want to investigate some of the standard decision-theoretic properties such as minimaxity and admissibility of the naive estimates and naive confidence sets of the means of the selected normal populations.

The problem of estimating the selected means can be connected to a model selection problem. Assume the usual linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\mathbf{X} \in R^{nxp}$ is a known matrix of variables and $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I})$. In model selection we want to choose a subset of variables of $\mathbf{X}$, which leads to estimating a subset of parameters $\boldsymbol{\beta}$. If we consider the case where $\mathbf{X} = \mathbf{I}$ and σ^2 is known, estimating $\boldsymbol{\beta}$'s corresponding to the largest values of $\mathbf{y}$ is equivalent to the selection problem we are considering.

1.1 Point Estimation

Assume $\mathbf{X} \sim N_p(\boldsymbol{\theta}, \sigma^2\mathbf{I})$. We are interested in estimating $\theta_{(1)} = \sum_{i=1}^{p} \theta_i I(X_i = X_{(1)})$ where $X_{(1)} \geq \ldots X_{(p)}$. For the problem of point estimation after selection the goal is to estimate the mean of the selected population. The naive estimator of $\theta_{(1)}$ is the sample maximum $X_{(1)}$ which will usually be positively biased, so some kind of shrinkage is required. Another related problem, known as ranking and selection, is to choose the best population, that is the population corresponding to the $\max\{\theta_1, \ldots, \theta_p\}$. In general, $\theta_{(1)}$ and $\max\{\theta_1, \ldots, \theta_p\}$ will not be the same.

Development of ranking and selection procedures started with the pioneering works of Bechhofer (1954) and Gupta (1956). Many extensive modifications and applications of this problems were considered since. For a comprehensive review of this research area see Gupta & Panchapakesan (1979), Gupta & Panchapakesan (1985) and Gibbons et al. (1999). A vast amount of research have been done for both ranking and selection problems, but we concentrate on the second aspect.

For the bivariate case with $p = 2$, that is, selecting one population out of two

populations, Stein (1964) proved that $X_{(1)}$ is admissible and minimax for estimating $\theta_{(1)}$. For the case $p \geq 2$ Stein proved that $X_{(1)}$ is biased and bias $\to \infty$ as $p \to \infty$, he also claims that $X_{(1)}$ will not be optimal in this case. Cohen & Sackrowitz (1982) provide an estimator for which the bias is much smaller than the bias of $X_{(1)}$.

Recently researchers have considered estimating the mean under certain sparsity assumptions. By sparsity we mean that most of elements of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ are equal to zero. Martin et al. (2014) used an empirical Bayes approach to get the posterior distribution for the sparse high-dimensional mean vector that concentrates on l_2 balls, centered at the true mean, with squared radius proportional to the minimax rate.

For the selection problem under sparsity, Reid & Tibshirani (2014) implicitly make an assumption of sparsity, with many effect sizes $\theta_i = 0$. We will discuss this paper in much more detail in Chapter 2. They adapted theory developed in Lee et al. (2014) where Reid & Tibshirani (2014) considered doing post-selection inference with the Lasso. Simon & Simon (2013) considered adjusting the selection bias of naive estimate from frequentist perspective. They start by estimating the mean by maximum likelihood, and then estimate the bias in order to achieve bias reduction. To further improve on this estimate Simon & Simon (2013) suggested estimating the second order bias as well. They also compare estimating bias for frequentist approach in their paper and Bayesian bias from empirical Bayes approach in Efron (2011), which turn out to be very similar. The advantage of Simon & Simon (2013) approach is that it is not limited to Gaussian setting.

1.2 Confidence intervals and confidence sets

1.2.1 Non-selection problem

Suppose $\mathbf{X} \sim N_p(\boldsymbol{\theta}, \sigma^2\mathbf{I})$ for $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$. When Stein (1956) discovered that $\mathbf{X}$ is an inadmissible estimate of $\boldsymbol{\theta}$ for $p \geq 3$, a lot of research was directed at finding dominating point estimates. The problem of confidence set estimation for multivariate normal mean is equally important, but it is much harder mathematically. Results in Brown (1966) implied that the usual confidence set is inadmissible for estimating mean for $p \geq 3$. Some papers deal with confidence sets under various losses, but most of the research is done to compare confidence sets with regard to shape, coverage probability, volume and conditional properties. Discussion of possible losses and conditional performance is given in Section 4.1.1 and Section 5.2.2

Substantial research has been devoted to constructing confidence intervals for a multivariate normal mean vector or on improving the usual one (Efron (2006); Casella & Hwang (1983)). Saxena & Tong (1969) gave a procedure for obtaining a fixed-width confidence interval for the largest mean of k normal populations.

Efron (2006) considered a general approach to constructing a confidence set with a volume smaller than that of the standard confidence set. He was also able to calculate a lower bound on the attainable volume using Bayesian and fiducial methods. He points out that reduced volume by itself does not guarantee improved inference.

Samworth (2005), following Casella & Hwang (1983), considered confidence sets recentered at the positive part James-Stein estimator of the form

$$C = \{\boldsymbol{\theta} : \|\delta_{JS}(\mathbf{X}) - \boldsymbol{\theta}\|^2 \leq w_\alpha(\boldsymbol{\theta})\}, \tag{1.1}$$

where $\delta_{JS}(\mathbf{X}) = \left(1 - \frac{a}{\|\mathbf{X}\|^2}\right)_+ \mathbf{X}$ is a positive part James-Stein estimator. Here, $\boldsymbol{\theta}$ is unknown and, thus, $w_\alpha(\boldsymbol{\theta})$ needs to be estimated. He considered two ways for

constructing radius function. First approach is based on approximating the upper α-point of the sampling distribution of $\|\delta_{JS}(\mathbf{X}) - \boldsymbol{\theta}\|^2$ using Taylor series expansion around origin. This set can be considerably smaller that the usual set, and is shown to dominate the usual set in terms of coverage probability for some $\boldsymbol{\theta}$. The second approach is based on parametric bootstrap and leads to a greater improvement in terms of volume.

For $p > 2$, Tseng et al. (1997) proposed confidence sets with constant coverage probability using "pseudo-empirical" Bayes approach. Here pseudo-empirical means that hyper parameters were estimated as functions of parameter $\boldsymbol{\theta}$, and not data $\mathbf{X}$ as it done in the empirical Bayes approach. These confidence sets were shown to have uniformly smaller volume than the usual confidence set. The disadvantage of their set is that it does not have an explicit form.

There is a substantial literature on the minimaxity and admissibility of the usual confidence set in non-selection context. In particular, Blumenthal (1970) considered the problem of interval estimation when the mean is restricted to be non-negative and the variance is known. He considered the following loss

$$L(\theta, I) = c\left(\delta_2(\cdot) - \delta_1(\cdot)\right) + I\left(\delta_1(\cdot) \leq \theta \leq \delta_2(\cdot)\right), \tag{1.2}$$

where the interval estimate $I(\cdot) = [\delta_1(\cdot), \delta_2(\cdot)]$ is closed with upper and lower end points δ_2 and δ_1, respectively. Under this loss, he proved that generalized Bayes interval under a uniform prior on $(0, \infty)$ is admissible. Morris (1983) was one of the first considered empirical Bayes confidence intervals that are shorter than standard confidence intervals.

Joshi (1969) first considered minimaxity and admissibility of the usual confidence sets for the mean of normal population. He proved that usual set is minimax and admissible for $p = 1$ and $p = 2$, but is not minimax and not admissible for $p \geq 3$. Brown (1966) and Joshi (1967) independently proved that recentering the usual

confidence set at a Stein type estimator leads to higher coverage probability for $p \geq 3$, although they did not provide an explicit form of that set.

Casella & Hwang (1991) considered the concepts of α-minimaxity and k-minimaxity (where k is the number of populations). They established that usual confidence set is both k-minimax and α-minimax under linear combination loss function which combines volume and coverage probability of the considered set.

He (1992) considered confidence interval estimation for the component of the normal mean. He proved inadmissibility of the classical confidence interval in terms of component loss using parametric empirical Bayes theory. Inadmissibility of the naive interval follows from domination of

$$C^b = \{\theta_1 : |\theta_1 - \delta^b(\mathbf{X})| \leq c\}, \tag{1.3}$$

where $\delta^b(\mathbf{X}) = \left(1 - \frac{b}{\|\mathbf{X}\|^2}\right)_+ X_1$, in terms of the coverage probability. He (1992) also investigated the construction of confidence intervals with variable radius based on Casella & Hwang (1983) and proved that resulting interval has smaller length than the naive intervals. In addition, he also provided numerical results that show that the constructed set dominates C_0 if $p \geq 3$.

Kabaila (2011) proved that for $p \geq 2$ the usual confidence interval for θ is admissible within a broad class of confidence intervals. In particular, their result establishes strong admissibility of the usual intervals, which complements results of Joshi (1969).

Most of the works mentioned above only deal with known variance case. If variance is unknown, the common strategy is to replace it by some estimator, such as a sample variance. There are few theoretical results for unknown variance.

1.2.2 Selection problems

Venter (1988) was one of the first to recognize that the naive intervals for the largest treatment mean have coverage probability below nominal level. He proposed some intervals that do maintain $1-\alpha$ confidence level. Venter (1988) also pointed out that it is enough for three or more θ_i's to be close to the largest for the actual coverage probability to drop below the nominal level. He treated upper and lower bounds separately, two-sided intervals are obtained by combining them. The resulting intervals are of the form

$$\{\theta_{(1)} : \big(X_{(1)} - z_{k,\alpha/2}, X_{(1)} + z_{\alpha/2}\big)\}, \tag{1.4}$$

where $z_{k,\alpha/2} = z_{1-(1-\alpha)^{1/k}}$, are not symmetric with upper bound is the same as for naive intervals and lower bound is smaller, since estimating $\theta_{(1)}$ by $X_{(1)}$ is biased upward. Building up on the idea of asymmetric confidence intervals of Venter (1988), Fuentes et al. (2014) derived intervals for selected means that guarantee coverage probability of $1-\alpha$, by working on the expression for the lower bound on coverage probability of the naive set. The resulting confidence intervals are centered at $X_{(1)}$ as the naive sets, but the length is a function of p and k and are chosen to guarantee good coverage probability.

Qiu & Hwang (2007) took an empirical Bayes approach to construct simultaneous confidence intervals with good coverage probability for the means of the selected populations. By generalizing the false discovery rate approach, Benjamini & Yekutieli (2005) proposed the false coverage-statement rate (FCR) as a measure of interval coverage following selection. Based on the concept of FCR, they provided confidence intervals that maintain $1-\alpha$ FCR. Building on FCR method, Zhao & Hwang (2012) proposed confidence intervals for pre-selected parameters that control Bayes FCR. Their approach leads to a shrinkage type confidence interval, which reduces selection bias and makes this intervals shorter compared to Benjamini & Yekutieli

(2005). These intervals perform especially well for sparse $\boldsymbol{\theta}$.

In addition to proposing improved point estimates for selection problem Reid & Tibshirani (2014) also proposed confidence intervals. Their intervals perform well, for example they dominate intervals proposed in Benjamini & Yekutieli (2005). Unfortunately, their intervals are numerically quite unstable, and we will not consider them in our subsequent analysis.

CHAPTER 2

POINT ESTIMATION

Consider a p-dimensional random vector $\mathbf{X} = (X_1, \ldots, X_p)$, which is normally distributed with unknown mean $\boldsymbol{\theta}$ and covariance matrix $\sigma^2\mathbf{I}$, that is $\mathbf{X} \sim N_p(\boldsymbol{\theta}, \sigma^2\mathbf{I})$. The variance σ^2 is known and assumed to be equal to one, unless stated otherwise. In general, the point estimates and confidence intervals will depend on n replications, but by the principle of sufficiency we can assume, without loss of generality, a single observation. Let the order of components be $X_{(1)} \geq X_{(2)} \geq \cdots \geq X_{(p)}$. We are interested in estimating the population mean corresponding to the largest sample mean, that is, we wish to estimate $\theta_{(1)} = \sum_{i=1}^{p} \theta_i I(X_i = X_{(1)})$. As was mentioned in the introduction the naive estimate, $X_{(1)}$, of $\theta_{(1)}$ is biased upwards. Before proceeding to construction of the improved estimates, we will investigate minimaxity and admissibility of $X_{(1)}$ as an estimate of $\theta_{(1)}$.

Problems like estimating the largest differential expression in genetics or largest differential activation of voxels in fMRI studies implicitly assume sparsity on the parameter space, that is a high proportion of parameters is assumed to be zero. We want to estimate those few non-zero parameters out of the large total number of parameters considered.

2.1 Minimaxity

Definition 1: An estimator δ^M of θ, evaluated using the risk function $R(\theta, \delta) = EL(\theta, \delta)$, is called a *minimax* estimator if it minimizes the maximum risk, that is, it satisfies

$$\inf_{\delta} \sup_{\theta} R(\theta, \delta) = \sup_{\theta} R(\theta, \delta^M). \tag{2.1}$$

The minimaxity question of the naive estimate of $\theta_{(1)}$ is settled. Here we provide

some of the results that were obtained in the literature. The most relevant result to the problem in consideration is that of Sackrowitz & Samuel-Cahn (1986). They proved that $X_{(1)}$ is minimax for estimating $\theta_{(1)}$ under squared error loss for $p = 2$, but is not minimax for $p > 2$. They prove the result by showing that the difference of supremums of risks of the estimator $\delta_0 = X_{(1)}$ and a general bias adjusted estimator $\delta_c = X_{(1)} - c$ is positive. The non-minimaxity of $X_{(1)}$ can be explained by the fact that $X_{(1)}$ is biased, and some bias-adjustment is required for a better estimate.

A number of papers give a different setup of the problem or a different minimaxity concept. Gupta & Miescke (1986) considered populations with different sample sizes and $p \geq 3$. They considered 0-1 loss and assumed variances are different but known. They proved that estimating the largest normal mean with the largest sample mean is minimax if the variances are equal. Their rule performs poorly whenever the parameters $\theta_1, \ldots, \theta_k$ are close together and does not seem to be a universally good decision rule. More research is needed to find rules when the naive estimate fails to be optimal. Gupta & Miescke (1986) also found that the naive estimate is uniformly best permutation invariant procedure if the sample sizes $n_1, \ldots, n_p$ are all equal.

Berger (1976c) considered an interesting concept of tail minimaxity. Stein type estimators outperform the usual estimator with particularly big gains around the origin. The hope is that such estimators, if they are also tail minimax, will have risk as small as that of $\mathbf{X}$ for large $\boldsymbol{\theta}$, and a reasonable behavior in the mid-range. Berger (1976c) used numerical results to support this. Tail minimaxity is a necessary condition for minimaxity which is easier to check than proving minimaxity, furthermore tail minimax estimators are often completely minimax.

Following Cohen & Sackrowitz (1982), Venter & Steel (1991) proposed a family of estimators, that are weighted averages of order statistics, for the selection problem. They also identify members of this family that are minimax for the selection

problem under squared error loss. Their estimates are very close to those of Cohen & Sackrowitz (1982), for that reason only comparisons with estimates of the latter are considered in simulations below.

2.2 Admissibility

Definition 2: An estimator δ, evaluated using the risk function $R(\theta, \delta)$, is said to be *inadmissible* if there exists another estimator δ' which dominates it, that is, such that

$$R(\theta, \delta') \leq R(\theta, \delta) \tag{2.2}$$

for all θ, with strict inequality for some θ, and is *admissible* if no such estimator δ' exists.

The most common ways to establish admissibility is to try to find a dominating estimator, and prove inadmissibility by definition or use Blyth method.

For **Blyth method** (Lehmann & Casella (2003)) suppose $\Theta \subset R^p$ and $R(\theta, \delta)$ is continuous in θ for all δ. Let δ be an estimator and π_n be a sequence of priors such that any open ball $B \subset \Theta$,

$$\frac{R(\pi_n, \delta) - R(\pi_n, \delta_{\pi_n})}{\pi_n(B)} \to 0 \text{ as } n \to \infty, \tag{2.3}$$

then δ is admissible.

The admissibility proof of $X_{(1)}$ for $p = 2$ has been credited to Brown (1979). He did prove some admissibility results for estimating $\max \theta_i$, which could have been confused with estimating $\theta_{(1)}$, which in general are not the same. Cohen & Strawderman (1973) consider a general problem of admissibility of the best invariant estimator of a location parameter for a very wide class of loss functions. They

provide sufficient conditions for admissibility which are based on the results form Brown (1966).

Berger (1976a) considered looking at coordinates of $\boldsymbol{\theta}$ separately. He proves that generalized Bayes estimator $\delta_F = X_1$, under the prior $F(\boldsymbol{\theta}) \equiv 1$, of θ_1 is admissible if $p \leq 3$ under some conditions. In a companion paper Berger (1976b) proved inadmissibility of the best invariant estimator X_1 of the first coordinate of $\boldsymbol{\theta}$ if $p \geq 4$.

Berger & Strawderman (1996) approach the problem of admissibility of normal means from a Bayesian perspective. They considered the following hierarchical model

$$\mathbf{X}|\boldsymbol{\theta}, \sigma^2 \sim N_p(\mathbf{z}\boldsymbol{\theta}, \sigma^2\mathbf{I}), \qquad (\boldsymbol{\theta}, \sigma^2) \sim \pi_1(\sigma^2)\pi_2(\boldsymbol{\theta}),$$

where $\mathbf{z}$ is a matrix of known covariates of rank k. They put a hierarchical prior on the hyper parameters $\boldsymbol{\theta}$, and quite general priors for σ^2 are considered $\pi_1(\sigma^2) \sim \frac{c}{(\sigma^2)^a}$, for some constants a and c. Sets of parameter configurations that give admissible and inadmissible mean estimators are provided. In particular, shrinkage priors for $\boldsymbol{\theta}$ are recommended. For example, common shrinkage prior $\pi_2(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|^{-(k-2)}$ (same prior that the one in Maruyama that gives admissibility, see Section 2.2.2) gives admissible mean estimator if $a \geq 1$ and inadmissible estimator if $a < 1$. Maruyama (2009) proved admissibility of the generalized Bayes estimators of the mean vector in a general class of spherically symmetric distribution. We will use Berger (1976a) and Maruyama (2009) as the basis for dealing with admissibility of $X_{(1)}$ as an estimator of $\theta_{(1)}$ in the selection context.

2.2.1 The Berger approach

Berger (1976a) considered the problem of estimating a coordinate of a location

vector and whether the resulting estimate will be admissible. In particular, he considers estimating θ_1, one of the coordinates of parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$. Here, we will adapt the argument for estimating $\theta_{(1)} = \sum_{i=1}^{p} \theta_i I(X_i = X_{(1)})$, the selected coordinate corresponding to the population with the largest sample mean. Denote the generalized Bayes estimator of $\theta_{(1)}$ with respect to prior G as δ_G.

Consider estimating $\theta_{(1)}$ under the squared error loss, $L(\delta, \boldsymbol{\theta}) = \left(\delta - \theta_{(1)}\right)^2$. The following conditions are needed.

Condition I:

i. All third order derivatives of G exist;

ii. $E_0 X_{(1)} = \int x_{(1)} p(\mathbf{x}|\boldsymbol{\theta})\, d\mathbf{x} = 0$;

iii. $\int p(\mathbf{x}|\boldsymbol{\theta}) G(\boldsymbol{\theta})\, d\boldsymbol{\theta} > 0$, for every $\mathbf{x} \in R^p$.

Under Condition I it follows that

$$\int L'(\delta_G - \theta_{(1)}) p(\mathbf{x}|\boldsymbol{\theta}) G(\boldsymbol{\theta})\, d\boldsymbol{\theta} = \tag{2.4}$$
$$= \sum_{i=1}^{p} \underbrace{\left[\int 2\left(\delta_G - \theta_i\right) p(\mathbf{x}|\boldsymbol{\theta}) G(\boldsymbol{\theta})\, d\boldsymbol{\theta}\right]}_{=0 \text{ since } \delta_G = \frac{\int \theta_i p(\mathbf{x}|\boldsymbol{\theta}) G(\boldsymbol{\theta})\, d\boldsymbol{\theta}}{\int p(\mathbf{x}|\boldsymbol{\theta}) G(\boldsymbol{\theta})\, d\boldsymbol{\theta}}} I(x_i = x_{(1)}) = 0.$$

Define the moment structure, as in Berger (1976a),

$$m_{j(1),j(2),\ldots,j(m)} = E_0\left[\left(\prod_{i=1}^{m} X_{j(i)}\right) L'(X_{(1)})\right] = \sum_{k=1}^{p} E_0\left[\left(\prod_{i=1}^{m} X_{j(i)}\right) X_{(k)}\right] I(X_k = X_{(1)}).$$

It is clear that for squared error loss $m_1 > 0$. Without loss of generality, reduce the moment structure of the problem following Berger (1976a), that is we assume $m_1 = 1$, and make a linear transformation $Y_{(1)} = X_{(1)}$, and $Y_i = X_i - m_i X_{(1)}$ for $i \geq 2$. For the transformed problem

$$m_1 = 1, \text{ and } m_i = 0 \text{ for } i \geq 2, \tag{2.5}$$

since

$$m_i = E_0\left[X_i L'(X_{(1)})\right] = 2E_0\left[X_i X_{(1)}\right] = 2\sum_{j=1}^{p} \underbrace{E_0\left[X_i X_j\right]}_{=0} I(X_j = X_{(1)}) = 0 \text{ for } i \geq 2.$$

We revert to X notation but assume that (2.5) holds. Let $M = \{m_{i,j}\}_{i,j=2}^{p} = E_0\left(X_i X_j X_{(1)}\right)$, $2 \leq i \leq p$, and $2 \leq j \leq p$. Since matrix M is symmetric we can find orthogonal matrix $P = \{p_{i,j}\}_{i,j=2}^{p}$, such that PMP^T is a diagonal matrix. Consider the final transformation $Y = XQ$, where Q is $p \times p$ matrix with elements $q_{1,1} = 1$, $q_{i,1} = q_{1,i} = 0$ for $2 \leq 0 \leq p$, and $q_{i,j} = p_{i,j}$ for $2 \leq i \leq p$ and $2 \leq j \leq p$.

For this transformed problem

$$m_1 = E_0\left[X_i X_{(1)}\right] = 1, \text{ and } m_i = 0 \text{ if } i \geq 2, \tag{2.6}$$

$$m_{i,j} = E_0\left[X_i X_j X_{(1)}\right] = \begin{cases} -1, \text{ if } 2 \leq i = j \leq r_1 + 1 \\ 1, \text{ if } r_1 + 2 \leq i = j \leq r_1 + r_2 + 1 \\ 0, \text{ otherwise, for } i \geq 2 \text{ and } j \geq 2. \end{cases} \tag{2.7}$$

Here we modify the proof in Berger (1976a) with respect to taking selection into account, the rest of the proof will be as in Berger (1976a).

The first step is to get approximation to $\gamma_G = \delta_G - x_{(1)} = \sum_{i=1}^{p}(\delta_G - x_i)I(x_i = x_{(1)})$. From (2.4) we get

$$2\sum_{i=1}^{p}\left[\int (\gamma_G + x_i - \theta_i)\, f(\mathbf{x} - \boldsymbol{\theta})G(\boldsymbol{\theta})\, d\boldsymbol{\theta}\right] I(x_i = x_{(1)}) = 0. \tag{2.8}$$

It then follows that

$$\delta_G(\mathbf{x}) = -\frac{\sum_{i=1}^{p}\left[\int (x_i - \theta_i) f(\mathbf{x} - \boldsymbol{\theta})G(\boldsymbol{\theta})\, d\boldsymbol{\theta}\right] I(x_i = x_{(1)})}{\int f(\mathbf{x} - \boldsymbol{\theta})G(\boldsymbol{\theta})\, d\boldsymbol{\theta}}.$$

Consider one term in the numerator and do a Taylor series expansion of $G(\boldsymbol{\theta})$

around $\boldsymbol{\theta} = \mathbf{x}$:

$$\begin{aligned}
\int (x_i - \theta_i) f(\mathbf{x} - \boldsymbol{\theta}) G(\boldsymbol{\theta})\, d\boldsymbol{\theta} &= \underbrace{\int (x_i - \theta_i) f(\mathbf{x} - \boldsymbol{\theta}) G(\mathbf{x})\, d\boldsymbol{\theta}}_{= G(\mathbf{x}) \int (x_i - \theta_i) f(\mathbf{x} - \boldsymbol{\theta})\, d\boldsymbol{\theta} = 0} \\
&\quad + \int (x_i - \theta_i) f(\mathbf{x} - \boldsymbol{\theta}) \sum_{j=1}^{p} \frac{\partial G}{\partial x_j}(\mathbf{x})(x_j - \theta_j)\, d\boldsymbol{\theta} \\
&\quad + \int (x_i - \theta_i) f(\mathbf{x} - \boldsymbol{\theta}) \frac{1}{2} \sum_{j,k=1}^{p} \frac{\partial^2 G}{\partial x_j \partial x_k}(\boldsymbol{x})(x_j - \theta_j)(x_k - \theta_k)\, d\boldsymbol{\theta} + \ldots \\
&\approx - \left[\frac{\partial G}{\partial x_i}(\boldsymbol{x}) + \frac{1}{2} \sum_{j,k \neq i}^{p} \frac{\partial^2 G}{\partial x_j \partial x_k}(\boldsymbol{x}) \right] \\
&\equiv \mathscr{D}_i^* G(\mathbf{x}). \qquad (2.9)
\end{aligned}$$

Similarly consider the denominator of δ_G

$$\begin{aligned}
\int f(\mathbf{x} - \boldsymbol{\theta}) G(\boldsymbol{\theta})\, d\boldsymbol{\theta} &= \int f(\mathbf{x} - \boldsymbol{\theta}) G(\mathbf{x})\, d\boldsymbol{\theta} \\
&\quad + \int f(\mathbf{x} - \boldsymbol{\theta}) \sum_{j=1}^{p} \frac{\partial G}{\partial x_j}(\mathbf{x})(x_j - \theta_j)\, d\boldsymbol{\theta} \\
&\quad + \int f(\mathbf{x} - \boldsymbol{\theta}) \frac{1}{2} \sum_{j,k \neq i}^{p} \frac{\partial^2 G}{\partial x_j \partial x_k}(\mathbf{x})(x_j - \theta_j)(x_k - \theta_k)\, d\boldsymbol{\theta} + \ldots \\
&\approx G(\mathbf{x}). \qquad (2.10)
\end{aligned}$$

Thus, we have an approximation for γ_G

$$\gamma_G(\mathbf{x}) \approx - \frac{\sum_{i=1}^{p} \mathscr{D}_i^* G(\mathbf{x}) I(x_i = x_{(1)})}{G(\mathbf{x})}. \qquad (2.11)$$

To prove admissibility of δ_G, we need to find a sequence of priors, $g_n(\boldsymbol{\theta})$, such that $\lim_{R \to \infty} g_n(\boldsymbol{\theta}) = 1$, and

$$\lim_{n \to \infty} \int \Delta_{\delta_R}^{G}(\theta_{(1)}) G(\boldsymbol{\theta}) g_n(\boldsymbol{\theta})\, d\boldsymbol{\theta} = 0. \qquad (2.12)$$

For simplicity consider $G(\boldsymbol{\theta}) \equiv 1$, then

$$\begin{aligned}\Delta_{\delta_n}^G &= R(\delta_G, \theta_{(1)}) - R(\delta_n, \theta_{(1)}) \\ &= \sum_{i=1}^{p} \int \left[(x_i - \theta_i)^2 - (\delta_n - \theta_i)^2\right] f(\mathbf{x} - \boldsymbol{\theta}) I(x_i = x_{(1)})\, d\mathbf{x}, \end{aligned} \tag{2.13}$$

where δ_n is a generalized Bayes rule with respect to prior $g_n(\boldsymbol{\theta})$.

Now expand the following around $x_i - \theta_i$

$$\begin{aligned}\sum_{i=1}^{p} (\delta_n - \theta_i)^2 I(x_i = x_{(1)}) &= \sum_{i=1}^{p} (\gamma_n + x_i - \theta_i)^2 I(x_i = x_{(1)}) \\ &= \sum_{i=1}^{p} \left[(x_i - \theta_i)^2 + 2(x_i - \theta_i)\gamma_n + \gamma_n^2\right] I(x_i = x_{(1)}),\end{aligned}$$

so that

$$\Delta_{\delta_n}^G = -\sum_{i=1}^{p} \int \left[2\gamma_n(x_i - \theta_i) + \gamma_n^2\right] f(\mathbf{x} - \boldsymbol{\theta}) I(x_i = x_{(1)})\, d\mathbf{x}. \tag{2.14}$$

Therefore, if we change the order of integration and make use of (2.9) and (2.10), we get the following expression for difference in the risks

$$\begin{aligned}&\int \Delta_{\delta_n}^G(\theta_{(1)}) g_n(\boldsymbol{\theta})\, d\boldsymbol{\theta} = \\ &\quad = -\int_\Theta \int_{\mathcal{X}} \sum_{i=1}^{p} \left[2\gamma_n(x_i - \theta_i) + \gamma_n^2\right] f(\mathbf{x} - \boldsymbol{\theta}) I(x_i = x_{(1)}) g_n(\boldsymbol{\theta})\, d\mathbf{x}\, d\boldsymbol{\theta} \\ &\quad = 2\int_{\mathcal{X}} \gamma_n \sum_{i=1}^{p} \left[\int_\Theta (x_i - \theta_i) f(\mathbf{x} - \boldsymbol{\theta}) g_n(\boldsymbol{\theta})\, d\boldsymbol{\theta}\right] I(x_i = x_{(1)})\, d\mathbf{x} \\ &\qquad - \int_{\mathcal{X}} \gamma_n^2 \sum_{i=1}^{p} \left[f(\mathbf{x} - \boldsymbol{\theta}) g_n(\boldsymbol{\theta})\, d\boldsymbol{\theta}\right] I(x_i = x_{(1)})\, d\mathbf{x} \\ &\quad = 2\int_{\mathcal{X}} \left[\frac{\sum_{i=1}^{p} \mathscr{D}_i^* g_n(\mathbf{x}) I(x_i = x_{(1)})}{g_n(\mathbf{x})}\right]^2 \sum_{i=1}^{p} g_n(\mathbf{x}) I(x_i = x_{(1)})\, d\mathbf{x} \\ &\qquad - \int_{\mathcal{X}} \frac{\left[\sum_{i=1}^{p} \mathscr{D}_i^* g_n(\mathbf{x}) I(x_i = x_{(1)})\right]^2}{g_n^2(\mathbf{x})} g_R(\mathbf{x})\, d\mathbf{x} \\ &\quad = \sum_{i=1}^{p} \int_{\mathcal{X}} \frac{\left[\mathscr{D}_i^* g_n(\mathbf{x})\right]^2}{g_n(\mathbf{x})} I(x_i = x_{(1)})\, d\mathbf{x}. \end{aligned} \tag{2.15}$$

Again, it seems that selection does not matter when we are working with the Bayes risks. After we integrate over $\mathbf{x}$, we will only have one non-zero component

in the sum corresponding to $X_{(1)}$. Selection only matters in the sense that we need to pick up the right component of $\boldsymbol{\theta}$ corresponding to the selected population and $X_{(1)}$. The only difference from the proof in Berger (1976a), is that the coordinate we are working with will be the selected coordinate $\theta_{(1)}$, as opposed to θ_1.

Berger (1976a) suggested the following sequence of priors that satisfies $\lim_{n\to\infty} g_n(\boldsymbol{\theta}) = 1$, and (2.15) above

$$g_n(\boldsymbol{\theta}) = \begin{cases} 1, & \text{if } \|\boldsymbol{\theta}\| \leq 1, \\ \left(1 - \frac{\ln \|\boldsymbol{\theta}\|}{\ln n}\right)^{23}, & \text{if } 1 \leq \|\boldsymbol{\theta}\| \leq n, \\ 0, & \text{if } \|\boldsymbol{\theta}\| > n, \end{cases} \tag{2.16}$$

where $\|\boldsymbol{\theta}\| = |\theta_{(1)}| + \sum_{i=1}^{p} \theta_i^2$. These priors can also be used in the selection context. To finally conclude that δ_G with respect to G is admissible for estimating $\theta_{(1)}$ we use the form of Stein's sufficient condition for admissibility given in Farrell (1964). As the result we have the following, which are similar to Corollary B1 and B2 in Berger (1976a).

Theorem 1. *Under the prior $G(\boldsymbol{\theta}) = (1 + \|\boldsymbol{\theta}\|^r)^{-1}$, where $r \geq 0$, the generalized Bayes estimator, $\delta_G(\mathbf{x}) = \sum_{i=1}^{p} \left[x_i - \frac{\mathscr{D}_i^* G(\mathbf{x})}{G(\mathbf{x})}\right] I(x_i = x_{(1)})$, is admissible for estimating $\theta_{(1)}$ if $p \leq r + 3$.*

Corollary 1. *For squared error loss, the best invariant estimator of $\theta_{(1)}$, $X_{(1)}$, is admissible if $p \leq 3$.*

Proof of Theorem 1 is based on the analog of Theorem B of Berger (1976a) under selection which states that under some technical assumptions on generalized prior $G(\boldsymbol{\theta})$, density $p(\mathbf{x}|\boldsymbol{\theta})$ and loss $L(\delta, \boldsymbol{\theta})$, generalized Bayes estimator, δ_G, of $\theta_{(1)}$ is admissible. Prior $G(\boldsymbol{\theta}) = (1 + \|\boldsymbol{\theta}\|^r)^{-1}$, squared error loss and multivariate normal density satisfy these conditions.

The admissibility proof of an estimate of a coordinate in Berger (1976a) is quite similar to the proof of the selected parameter. The only difference is that every term which is a function of selected sample mean, that is $X_{(1)}$ and $\theta_{(1)}$ will have the following form $\sum_{i=1}^{p} s(X_i)I(X_i = X_{(1)})$, where s is some function. But since only one term corresponding to $s(X_{(1)})$ is non-zero in the sum, it will not affect the proof. We just need to replace X_1 and θ_1 in the proof by $X_{(1)}$ and $\theta_{(1)}$.

2.2.2 The Maruyama approach

Case $k = 1$

Based on the proof technique of Brown & Hwang (1982) for deducing admissibility, Maruyama (2009) developed a method for giving sufficient conditions for the admissibility of generalized Bayes rule. He used a certain adaptive sequence of proper priors that approaches improper harmonic prior fast enough to establish admissibility for a general class of spherically symmetric distributions. We will use similar approach for proving admissibility in the selection case.

To use Blyth method and prove the admissibility of the generalized Bayes estimator under the prior g, we need to show that for a sequence of priors satisfying $\int_{\|\boldsymbol{\theta}\|\geq 1} g_n(\boldsymbol{\theta})\, d\boldsymbol{\theta} > c$ for some positive c, difference of risks of generalized Bayes estimators under g and g_n converges to zero. We will consider the following priors proposed in Maruyama (2009). Consider the sequence of priors

$$g_n(\boldsymbol{\theta}) = g(\boldsymbol{\theta})h_n^2(\boldsymbol{\theta}) = G(\|\boldsymbol{\theta}\|)H_n^2(\|\boldsymbol{\theta}\|), \tag{2.17}$$

where

$$H_n(\eta) = \frac{\int_\eta^\infty e^{(\eta-r)/n}\beta(r)\, dr}{\int_\eta^\infty \beta(r)\, dr}, \quad n = 1, 2, \ldots, \tag{2.18}$$

$$\beta(r) = -\frac{d}{dr}\left\{\left(\int_1^{2+r} \frac{s^{1-p}}{G(s)}\, ds\right)^{-1}\right\} = \frac{(r+2)^{1-p}/G(2+r)}{\left(\int_1^{2+r} \{s^{1-p}/G(s)\}\ ds\right)^2}, \tag{2.19}$$

and G is assumed to be regularly varying. Various properties of $\beta(\cdot)$ and $H_n(\cdot)$ are given in Theorems 2.1-2.3 in Maruyama (2009).

Maruyama (2009) also needs the following regularity condition

$$\int_1^{\infty} \frac{r^{1-p}}{G(r)}\, dr = \infty. \tag{2.20}$$

To proceed with admissibility proof, first, we need to define a generalized Bayes estimate of $\theta_{(1)}$, δ_g, under the prior $g(\boldsymbol{\theta})$

$$\begin{aligned}
\delta_g(\mathbf{x}) &= \frac{\int \theta_{(1)} p(\mathbf{x}|\boldsymbol{\theta}) g(\boldsymbol{\theta})\, d\boldsymbol{\theta}}{\int p(\mathbf{x}|\boldsymbol{\theta}) g(\boldsymbol{\theta})\, d\boldsymbol{\theta}} \\
&= x_{(1)} + \frac{\int (\theta_{(1)} - x_{(1)}) p(\mathbf{x}|\boldsymbol{\theta}) g(\boldsymbol{\theta})\, d\boldsymbol{\theta}}{\int p(\mathbf{x}|\boldsymbol{\theta}) g(\boldsymbol{\theta})\, d\boldsymbol{\theta}} \\
&= x_{(1)} + \frac{\int \sum_{i=1}^{p} \frac{\partial g}{\partial \theta_i} p(\mathbf{x}|\boldsymbol{\theta}) I(x_i = x_{(1)})\, d\boldsymbol{\theta}}{\int p(\mathbf{x}|\boldsymbol{\theta}) g(\boldsymbol{\theta})\, d\boldsymbol{\theta}} \\
&= \sum_{i=1}^{p} \left[x_i + \frac{\int \frac{\partial g}{\partial \theta_i} p(\mathbf{x}|\boldsymbol{\theta})\, d\boldsymbol{\theta}}{\int p(\mathbf{x}|\boldsymbol{\theta}) g(\boldsymbol{\theta})\, d\boldsymbol{\theta}} \right] I(x_i = x_{(1)}).
\end{aligned} \tag{2.21}$$

Consider the difference in the Bayes risks of δ_g and δ_{g_n} with respect to the density g_n defined in (2.17), and a spherically symmetric target prior density $g(\boldsymbol{\theta})$. Then the integrated risk difference is

$$\begin{aligned}
\Delta_n &= \int \left[R(\boldsymbol{\theta}, \delta_g) - R(\boldsymbol{\theta}, \delta_{g_n})\right] g(\boldsymbol{\theta}) h_n^2(\boldsymbol{\theta})\, d\boldsymbol{\theta} \\
&= \int_{\mathcal{X}} (\delta_g - \delta_{g_n})^2 \int_{\Theta} p(\mathbf{x}|\boldsymbol{\theta}) g(\boldsymbol{\theta}) h_n^2(\boldsymbol{\theta})\, d\boldsymbol{\theta} \\
&= \int_{\mathcal{X}} \left(\sum_{i=1}^{p} \left\{ \frac{\int \frac{\partial g}{\partial \theta_i} p(\mathbf{x}|\boldsymbol{\theta})\, d\boldsymbol{\theta}}{\int g(\boldsymbol{\theta}) p(\mathbf{x}|\boldsymbol{\theta})\, d\boldsymbol{\theta}} - \frac{\int \frac{\partial g_n}{\partial \theta_i} p(\mathbf{x}|\boldsymbol{\theta})\, d\boldsymbol{\theta}}{\int g(\boldsymbol{\theta}) p(\mathbf{x}|\boldsymbol{\theta})\, d\boldsymbol{\theta}} \right\} I(x_i = x_{(1)}) \right)^2 \\
&\qquad \times \int_{\Theta} p(\mathbf{x}|\boldsymbol{\theta}) g(\boldsymbol{\theta}) h_n^2(\boldsymbol{\theta})\, d\boldsymbol{\theta}\, d\mathbf{x}
\end{aligned}$$

$$
\begin{aligned}
&= \int_{\mathcal{X}} \left(\sum_{i=1}^{p} \left\{ \frac{\int \frac{\partial g}{\partial \theta_i} p(\mathbf{x}|\boldsymbol{\theta})\, d\boldsymbol{\theta}}{\int g(\boldsymbol{\theta}) p(\mathbf{x}|\boldsymbol{\theta})\, d\boldsymbol{\theta}} - \frac{\int \frac{\partial g}{\partial \theta_i} h_n^2 p(\mathbf{x}|\boldsymbol{\theta})\, d\boldsymbol{\theta}}{\int g(\boldsymbol{\theta}) h_n^2 p(\mathbf{x}|\boldsymbol{\theta})\, d\boldsymbol{\theta}} - \frac{\int 2h_n \frac{\partial h_n}{\partial \theta_i} g(\boldsymbol{\theta}) p(\mathbf{x}|\boldsymbol{\theta})\, d\boldsymbol{\theta}}{\int g(\boldsymbol{\theta}) h_n^2 p(\mathbf{x}|\boldsymbol{\theta})\, d\boldsymbol{\theta}} \right\} \right. \\
&\qquad \left. \times I(x_i = x_{(1)}) \right)^2 \int_{\Theta} p(\mathbf{x}|\boldsymbol{\theta}) g(\boldsymbol{\theta}) h_n^2(\boldsymbol{\theta})\, d\boldsymbol{\theta}\, d\mathbf{x} \\
&\leq 2 \int_{\mathcal{X}} \left(\sum_{i=1}^{p} \frac{\int 2h_n \frac{\partial h_n}{\partial \theta_i} g(\boldsymbol{\theta}) p(\mathbf{x}|\boldsymbol{\theta})\, d\boldsymbol{\theta}}{\int g(\boldsymbol{\theta}) h_n^2 p(\mathbf{x}|\boldsymbol{\theta})\, d\boldsymbol{\theta}} I(x_i = x_{(1)}) \right)^2 \int_{\Theta} p(\mathbf{x}|\boldsymbol{\theta}) g(\boldsymbol{\theta}) h_n^2(\boldsymbol{\theta})\, d\boldsymbol{\theta}\, d\mathbf{x} \\
&\qquad + 2 \int_{\mathcal{X}} \left(\sum_{i=1}^{p} \left\{ \frac{\int \frac{\partial g}{\partial \theta_i} g(\boldsymbol{\theta}) p(\mathbf{x}|\boldsymbol{\theta})\, d\boldsymbol{\theta}}{\int g(\boldsymbol{\theta}) p(\mathbf{x}|\boldsymbol{\theta})\, d\boldsymbol{\theta}} - \frac{\int \frac{\partial g}{\partial \theta_i} h_n^2 p(\mathbf{x}|\boldsymbol{\theta})\, d\boldsymbol{\theta}}{\int g(\boldsymbol{\theta}) h_n^2 p(\mathbf{x}|\boldsymbol{\theta})\, d\boldsymbol{\theta}} \right\} I(x_i = x_{(1)}) \right)^2 \\
&\qquad \times \int_{\Theta} p(\mathbf{x}|\boldsymbol{\theta}) g(\boldsymbol{\theta}) h_n^2(\boldsymbol{\theta})\, d\boldsymbol{\theta}\, d\mathbf{x} \\
&= A_n + B_n. \qquad (2.22)
\end{aligned}
$$

Consider the first term in the sum and apply Cauchy-Schwartz inequality. It follows that

$$
\begin{aligned}
A_n &= 2 \int_{\mathcal{X}} \left(\sum_{i=1}^{p} \frac{\int 2h_n \frac{\partial h_n}{\partial \theta_i} g(\boldsymbol{\theta}) p(\mathbf{x}|\boldsymbol{\theta})\, d\boldsymbol{\theta}}{\int g(\boldsymbol{\theta}) h_n^2 p(\mathbf{x}|\boldsymbol{\theta})\, d\boldsymbol{\theta}} I(x_i = x_{(1)}) \right)^2 \int_{\Theta} p(\mathbf{x}|\boldsymbol{\theta}) g h_n^2\, d\boldsymbol{\theta}\, d\mathbf{x} \\
&\leq 2 \int_{\mathcal{X}} \left(\sum_{i=1}^{p} \frac{\int 2h_n \frac{\partial h_n}{\partial \theta_i} g(\boldsymbol{\theta}) p(\mathbf{x}|\boldsymbol{\theta})\, d\boldsymbol{\theta}}{\int g(\boldsymbol{\theta}) h_n^2 p(\mathbf{x}|\boldsymbol{\theta})\, d\boldsymbol{\theta}} \right)^2 \int_{\Theta} p(\mathbf{x}|\boldsymbol{\theta}) g h_n^2\, d\boldsymbol{\theta}\, d\mathbf{x} \\
&= 8p \int_{\mathcal{X}} \left(\int_{\Theta} g h_n \frac{\partial h_n}{\partial \theta_{(1)}} p(\mathbf{x}|\boldsymbol{\theta})\, d\boldsymbol{\theta} \right)^2 \frac{1}{\int_{\Theta} g h_n^2 p(\mathbf{x}|\boldsymbol{\theta})\, d\boldsymbol{\theta}}\, d\mathbf{x} \\
&\leq 8p \int_{\mathcal{X}} \frac{\int g h_n^2 p(\mathbf{x}|\boldsymbol{\theta})\, d\boldsymbol{\theta} \int g \left(\frac{\partial h_n}{\partial \theta_{(1)}} \right)^2 p(\mathbf{x}|\boldsymbol{\theta})\, d\boldsymbol{\theta}}{\int g h_n^2 p(\mathbf{x}|\boldsymbol{\theta})\, d\boldsymbol{\theta}}\, d\mathbf{x} \\
&= 8p \int_{\mathcal{X}} \int_{\Theta} g \left(\frac{\partial h_n}{\partial \theta_{(1)}} \right)^2 p(\mathbf{x}|\boldsymbol{\theta})\, d\boldsymbol{\theta}\, d\mathbf{x} \\
&= 8p \int g(\boldsymbol{\theta}) \left(\frac{\partial h_n}{\partial \theta_{(1)}} \right)^2 d\boldsymbol{\theta}. \qquad (2.23)
\end{aligned}
$$

Now by a p-spherical transformation

$$
\begin{aligned}
\theta_{(1)} &= t \cos \phi_1, \\
\theta_{(2)} &= t \sin \phi_1 \cos \phi_2, \qquad (2.24) \\
&\cdots \\
\theta_{(p)} &= t \sin \phi_1 \ldots \sin \phi_{p-2} \sin \phi_{p-1}.
\end{aligned}
$$

It follows that as $n \to \infty$

$$A_n \le 8p \int_0^{2\pi} \dots \int_0^{\pi} \int_0^{\infty} g(t) \left(\frac{dh_n(t)}{dt} \right)^2 t^{p-1} \frac{1}{\cos^2 \phi_{(1)}} \tag{2.25}$$

$$\times \sin^{p-2} \phi_{(1)} \sin^{p-3} \phi_{(2)} \dots \sin \phi_{(p-2)} \, dt \, d\phi_{(1)} \dots \, d\phi_{(p-1)}$$

$$= 8p \int_0^{\infty} g(t) \left(\frac{dh_n(t)}{dt} \right)^2 t^{p-1} \underbrace{\int_0^{\pi} \left[\frac{\sin^{p-2} \phi_{(1)}}{\cos^2 \phi_{(1)}} \right] d\phi_{(1)}}_{<\infty \text{ for } p \ge 4} \, dt \to 0. \tag{2.26}$$

Similarly, for the expression under the second integral in B_n as $n \to \infty$ we get

$$\left(\sum_{i=1}^{p} \left\{ \frac{\int \frac{\partial g}{\partial \theta_i} g(\boldsymbol{\theta}) p(\mathbf{x}|\boldsymbol{\theta}) \, d\boldsymbol{\theta}}{\int g(\boldsymbol{\theta}) p(\mathbf{x}|\boldsymbol{\theta}) \, d\boldsymbol{\theta}} - \frac{\int \frac{\partial g}{\partial \theta_i} h_n^2 p(\mathbf{x}|\boldsymbol{\theta}) \, d\boldsymbol{\theta}}{\int g(\boldsymbol{\theta}) h_n^2 p(\mathbf{x}|\boldsymbol{\theta}) \, d\boldsymbol{\theta}} \right\} I(x_i = x_{(1)}) \right)^2$$

$$\le \sum_{i=1}^{p} \left(\frac{\int \frac{\partial g}{\partial \theta_i} g(\boldsymbol{\theta}) p(\mathbf{x}|\boldsymbol{\theta}) \, d\boldsymbol{\theta}}{\int g(\boldsymbol{\theta}) p(\mathbf{x}|\boldsymbol{\theta}) \, d\boldsymbol{\theta}} - \frac{\int \frac{\partial g}{\partial \theta_i} h_n^2 p(\mathbf{x}|\boldsymbol{\theta}) \, d\boldsymbol{\theta}}{\int g(\boldsymbol{\theta}) h_n^2 p(\mathbf{x}|\boldsymbol{\theta}) \, d\boldsymbol{\theta}} \right)^2 \to 0,$$

which was shown in Maruyama (2009).

Theorem 2. *The generalized Bayes estimator*

$$\delta_G(\mathbf{x}) = x_{(1)} - \frac{\int (x_{(1)} - \theta_{(1)}) p(\mathbf{x}|\boldsymbol{\theta}) \frac{\partial G}{\partial \theta_{(1)}} \, d\boldsymbol{\theta}}{p(\mathbf{x}|\boldsymbol{\theta}) G(\boldsymbol{\theta}) \, d\boldsymbol{\theta}} \tag{2.27}$$

of $\theta_{(1)}$ with respect to $G(\|\boldsymbol{\theta}\|)$ is admissible if $p \ge 4$ under the regularity condition (2.20).

In particular,

Corollary 2. *The generalized Bayes estimator, $\delta_G(\mathbf{x})$, of $\theta_{(1)}$ with respect to the harmonic prior $G(\|\boldsymbol{\theta}\|) = \|\boldsymbol{\theta}\|^{2-p}$ or prior $G(\|\boldsymbol{\theta}\|) = \|\boldsymbol{\theta}\|^{2-p} \log(\|\boldsymbol{\theta}\| + c)$, $c > 1$, for $p \ge 4$ is admissible.*

Note that using this result we cannot prove admissibility of $X_{(1)}$, since in this case for $X_{(1)}$ to be generalized Bayes estimate of $\theta_{(1)}$, we should have $G(\|\boldsymbol{\theta}\|) \equiv 1$, but in this case the integral in the result above will be finite, and we cannot conclude admissibility. The best invariant estimator $X_{(1)}$ will be biased upward for estimating

$\theta_{(1)}$, and needs to be adjusted to remove this bias, which is exactly what generalized Bayes estimate for estimating the selected mean is doing.

For the case when $r = p - 2$ in Theorem 1, Theorem 1 and Theorem 3 coincide and give admissibility of the generalized Bayes rule under the prior $G(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|^r$.

From Corollary 1 we get admissibility of $X_{(1)}$ for estimating $\theta_{(1)}$ for $p \leq 3$, and from Corollary 2 generalized Bayes estimate of $\delta_G(x)$ of (2.27) is admissible for $p \geq 4$.

General k

For the case $k > 1$, we are interested in estimating k selected paramaters $\boldsymbol{\theta}_s = (\theta_{(1)}, \ldots, \theta_{(k)})$. The generalized Bayes estimate of $\boldsymbol{\theta}_s$ under the prior $g(\boldsymbol{\theta})$ is

$$\begin{aligned}\delta_g(\mathbf{x}) &= \frac{\boldsymbol{\theta}_s p(\mathbf{x}|\boldsymbol{\theta}) g(\boldsymbol{\theta})\, d\boldsymbol{\theta}}{\int p(\mathbf{x}|\boldsymbol{\theta}) g(\boldsymbol{\theta})\, d\boldsymbol{\theta}} \\ &= \sum_{i=1}^{p} \sum_{j=1}^{k} \left[x_i + \frac{\int (\theta_i - x_i) p(\mathbf{x}|\boldsymbol{\theta}) \nabla_k g\, d\boldsymbol{\theta}}{\int p(\mathbf{x}|\boldsymbol{\theta}) g(\boldsymbol{\theta})\, d\boldsymbol{\theta}} \right] I\left(x_i = x_{(j)}\right), \end{aligned} \tag{2.28}$$

where $\nabla_k g$ denotes the following vector of partial derivatives $\left(\frac{\partial g(\boldsymbol{\theta})}{\partial \theta_{(1)}}, \ldots, \frac{\partial g(\boldsymbol{\theta})}{\partial \theta_{(k)}}\right)$.

The beginning of the proof is the same as for the case $k = 1$ in the previous section, except that instead of $\frac{\partial g}{\partial \theta_{(1)}}$ we will have $\nabla_k g$. The term A_n will be

$$A_n \leq 4L_1 \int g(\|\boldsymbol{\theta}\|) \left\|\nabla_k h_n(\|\boldsymbol{\theta}\|)\right\|^2 d\boldsymbol{\theta}, \tag{2.29}$$

where L_1 is some constant.

Again make a k-spherical transformation as in (2.24), then

$$\frac{\partial h_n(t)}{\partial \theta_{(1)}} = \frac{1}{\cos\phi_1}\frac{\partial h(t)}{\partial t},$$
$$\frac{\partial h_n(t)}{\partial \theta_{(2)}} = \frac{1}{\sin\phi_1\cos\phi_2}\frac{\partial h(t)}{\partial t},$$
$$\dots$$
$$\frac{\partial h_n(t)}{\partial \theta_{(k)}} = \frac{1}{\sin\phi_1\dots\sin\phi_{k-1}\cos\phi_k}\frac{\partial h(t)}{\partial t}.$$

Applying this transformation, we get

$$A_n \le 4L_1\int_0^\infty\int_0^{2\pi}\int_0^\pi\dots\int_0^\pi g(t)\left(\frac{dh(t)}{dt}\right)^2\left[\frac{1}{\cos\phi_1}+\frac{1}{\sin\phi_1\cos\phi_2}+\dots\right.$$
$$\left.+\frac{1}{\sin\phi_1\dots\sin\phi_{k-1}\cos\phi_k}\right]t^{p-1}\sin^{p-2}\phi_1\sin^{p-3}\phi_2\dots\sin\phi_{p-2}\,d\phi_1\dots\,d\phi_{p-1}\,dt.$$

Consider separately integrals that involve trigonometric functions

$$\int_0^{2\pi}\int_0^\pi\dots\int_0^\pi\frac{\sin^{p-2}\phi_1}{\cos^2\phi_1}\sin^{p-3}\phi_2\dots\sin\phi_{p-2}\,d\phi_1\dots\,d\phi_{p-1}\text{ is finite for }p\ge 4,$$
$$\int_0^{2\pi}\int_0^\pi\dots\int_0^\pi\sin^{p-4}\phi_1\frac{\sin^{p-3}\phi_2}{\cos^2\phi_2}\dots\sin\phi_{p-2}\,d\phi_1\dots\,d\phi_{p-1}\text{ is finite for }p\ge 5,$$
$$\dots$$
$$\int_0^{2\pi}\int_0^\pi\dots\int_0^\pi\sin^{p-4}\phi_1\sin^{p-5}\phi_2\dots\frac{\sin^{p-(k+1)}\phi_k}{\cos^2\phi_k}\,d\phi_1\dots\,d\phi_{p-1}\text{ is finite for }p\ge k+3.$$

Thus as $n\to\infty$ we have

$$A_n \le L_2\int_0^\infty g(t)\left(\frac{dh(t)}{dt}\right)t^{p-1}\,dt\to 0 \text{ if } p\ge k+3. \tag{2.30}$$

The proof of $B_n\to 0$ is exactly the same as in Maruyama (2009), since in this paper the norm is a sum of p terms corresponding to partial derivatives with respect to $\theta_{(1)},\dots,\theta_{(p)}$ was shown to go to zero by showing that limit of each term in the sum is zero. Here, everything is the same except now we only have k terms corresponding to partial derivatives with respect to $\theta_{(1)},\dots,\theta_{(k)}$ which were shown to go to zero.

Theorem 3. *The generalized Bayes estimator of* $\boldsymbol{\theta}_s$ *with respect to* $G(\|\boldsymbol{\theta}\|)$ *is admissible if* $p \geq k+3$ *under the regularity condition* (2.20).

2.3 Examples

In this section we will get analytical expressions for generalized Bayes estimates (2.21) under harmonic and horseshoe priors. We also carry out a numerical study of the risk properties of these two Bayes estimates.

2.3.1 Generalized Bayes rules under harmonic prior

In the previous section we proved that the generalized Bayes estimator of $\theta_{(1)}$ is admissible under the prior $g(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|^{2-p}$. Under this prior, estimator has the following form

$$\delta_g(\mathbf{x}) = \left(\frac{1}{2\pi}\right)^{p/2} \sum_{i=1}^{p} \left[x_i + \frac{\int \cdots \int (x_i - \theta_i) \|\boldsymbol{\theta}\|^{2-p} e^{-\frac{\|\mathbf{x}-\boldsymbol{\theta}\|^2}{2}} \, d\boldsymbol{\theta}}{\int \|\boldsymbol{\theta}\|^{2-p} e^{-\frac{\|\mathbf{x}-\boldsymbol{\theta}\|^2}{2}} \, d\boldsymbol{\theta}} \right] I(x_i = x_{(1)}). \quad (2.31)$$

Consider the integral in numerator, and make transformation $\sum_{j\neq i}^{p} \theta_j = v$,

$$\begin{aligned} &\int (x_i - \theta_i) \|\boldsymbol{\theta}\|^{2-p} e^{-\frac{\|\mathbf{x}-\boldsymbol{\theta}\|^2}{2}} \, d\boldsymbol{\theta} = \\ &\qquad = \int (x_i - \theta_i) e^{-\frac{(x_i-\theta_i)^2}{2}} \left(\theta_i^2 + \sum_{j\neq i}^{p} \theta_j^2 \right)^{1-p/2} e^{-\frac{\sum_{j\neq i}^{p} (x_j-\theta_j)^2}{2}} \, d\boldsymbol{\theta} \\ &\qquad = \int_{-\infty}^{\infty} (x_i - \theta_i) e^{-\frac{(x_i-\theta_i)^2}{2}} \left[\int_0^{\infty} \left(\theta_i^2 + v\right)^{1-p/2} f(v) \, dv \right] d\theta_i, \end{aligned}$$

here θ_i and v are independent, and $f(v)$ is a noncentral chi-square density function with dimension $p-1$ and non-centrality parameter $\sum_{j\neq i}^{p} x_j^2$. Now integral in numerator is only two-dimensional and can be estimated numerically.

From Xu (2007) we know that denominator has a closed form (the form depends on whether p is even or odd). Combining these results we can write the form of generalized Bayes estimate of θ_1 under harmonic prior (this is for even $p \geq 4$)

$$\delta_g(\mathbf{x}) = \sum_{i=1}^{p} \left[x_i - \frac{\int_{-\infty}^{\infty} (x_i - \theta_i) e^{-\frac{(x_i-\theta_i)^2}{2}} \left[\int_0^{\infty} (\theta_i^2 + v)^{1-p/2} f(v)\, dv \right] d\theta_i}{\|\mathbf{x}\|^{2-p} \left(1 - e^{-\frac{\|\mathbf{x}\|^2}{2}} \sum_{m=0}^{p/2-2} \frac{(\|\mathbf{x}\|^2/2)^m}{m!} \right)} \right] I(x_i = x_{(1)}). \tag{2.32}$$

For odd $p \geq 5$, the denominator has the following form

$$\int \|\boldsymbol{\theta}\|^{2-p} e^{-\frac{\|\mathbf{x}-\boldsymbol{\theta}\|^2}{2}}\, d\boldsymbol{\theta} = \|\mathbf{x}\|^{2-p} \left[2\Phi(\|\mathbf{x}\|) - 1 - \sqrt{\frac{2}{\pi}} e^{-\frac{\|\mathbf{x}\|^2}{2}} \sum_{m=0}^{\frac{p-1}{2}-2} \frac{\|\mathbf{x}\|^{2m+1}}{(2m+1)!!} \right], \tag{2.33}$$

where $n!! = \prod_{k=0}^{m}(n-2k) = n(n-2)(n-4)\ldots$.

Let's try to get a closed form expression for the integral in the numerator of $\delta_g(\mathbf{x})$. Denote $k = p-1$, $\lambda = \sum_{j=2}^{p} x_i^2$, and $n = \frac{p}{2} - 1$, and consider the following integral

$$\begin{aligned}
\int_0^{\infty} \left(\theta_i^2 + v\right)^{1-p/2} f(v)\, dv &= E\left[a + \chi^2_{(k,\lambda)}\right]^{-n} \\
&= e^{-\lambda/2} \sum_{j=0}^{\infty} \frac{(\lambda/2)^j}{j!} E\left[\theta_i^2 + \chi^2_{k+2j}\right]^{-n} \\
&= e^{-\lambda/2} \sum_{j=0}^{\infty} \frac{(\lambda/2)^j}{j!} \int_0^{\infty} (\theta_i^2 + v)^{-n} \frac{1}{2^{\frac{k+2j}{2}} \Gamma\left(\frac{k+2j}{2}\right)} v^{\frac{k+2j}{2}-1} e^{-v/2}\, dv \\
&= e^{-\lambda/2} \sum_{j=0}^{\infty} \frac{(\lambda/2)^j}{j!} \frac{1}{2^{\frac{k+2j}{2}} \Gamma\left(\frac{k+2j}{2}\right)} \int_0^{\infty} \left(\theta_i^2 + v\right)^{-n} v^{\frac{k+2j}{2}-1} e^{-v/2}\, dv \\
&= e^{-\lambda/2} \sum_{j=0}^{\infty} \frac{(\lambda/2)^j}{j!} \frac{1}{2^{\frac{k+2j}{2}} \Gamma\left(\frac{k+2j}{2}\right)} \Gamma\left(\frac{k+2j}{2}\right) (\theta_i^2)^{\frac{k+2j}{2}-n} \\
&\qquad \times \Psi\left(\frac{k+2j}{2}, \frac{k+2j}{2} + 1 - n; \frac{\theta_i^2}{2}\right),
\end{aligned}$$

where $\Psi(\alpha, c; z) = \frac{\Gamma(1-c)}{\Gamma(1+a-c)} {}_1F_1(a; c; z) + \frac{\Gamma(c-1)}{\Gamma(a)} z^{1-c} {}_1F_1(1+a-c; 2-c; z)$ is Tricomi's

confluent hypergeometric function. Here

$$
\begin{aligned}
{}_1F_1(a;c;z) &= \sum_{l=0}^{\infty} \frac{(a)_l z^l}{(c)_l l!}, \\
{}_2F_1(a,b;c,z) &= \sum_{l=0}^{\infty} \frac{(a)_l (b)_l z^l}{(c)_l l!}, \text{ for } |z|<1.
\end{aligned}
$$

Thus we get the following expression in the numerator of $\delta_g(\mathbf{x})$, by making transformation $\theta_i^2 = y \sim \chi_1^2(x_i^2)$,

$$
\begin{aligned}
N \equiv & \int_{-\infty}^{\infty} e^{-\lambda/2} \sum_{j=0}^{\infty} \frac{(\lambda/2)^j}{j!} 2^{-\frac{p-1+2j}{2}} \theta_i^{2j+1} \Psi\left(\frac{p-1+2j}{2}, \frac{2j+3}{2}; \frac{\theta_i^2}{2}\right) (x_i-\theta_i) e^{-\frac{(x_i-\theta_i)^2}{2}} d\theta_i \\
= & e^{-\lambda/2} \sum_{j=0}^{\infty} \frac{(\lambda/2)^j}{j!} 2^{-\frac{p-1+2j}{2}} E\left\{\left(x_i y^{\frac{2j+1}{2}} - y^{j+1}\right) \Psi\left(\frac{p-1+2j}{2}, \frac{2j+3}{2}; \frac{y}{2}\right)\right\} dy \\
= & e^{-\lambda/2} \sum_{j=0}^{\infty} \frac{(\lambda/2)^j}{j!} 2^{-\frac{p-1+2j}{2}} \left[\sum_{m=0}^{\infty} e^{-x_i/2} \frac{(x_i/2)^m}{m!} \right. \\
& \left. \times E\left\{\left(x_i u^{j+\frac{1}{2}} - u^j\right) \Psi\left(\frac{p-1+2j}{2}, \frac{2j+3}{2}; \frac{u}{2}\right)\right\}\right]
\end{aligned}
$$

where $u \sim \chi_{1+2m}^2$. Then the above is equal to

$$
\begin{aligned}
N = & e^{-\lambda/2} \sum_{j=0}^{\infty} \frac{(\lambda/2)^j}{j!} 2^{-\frac{p-1+2j}{2}} \left[\sum_{m=0}^{\infty} e^{-x_i/2} \frac{(x_i/2)^m}{m!} \right. \\
& \left. \times \int_0^{\infty} \frac{1}{\Gamma\left(\frac{1+2m}{2}\right) 2^{\frac{1+2m}{2}}} u^{\frac{1+2m}{2}-1} e^{-u/2} \left(x_i u^{j+\frac{1}{2}} + u^{j+1}\right) \Psi\left(\frac{p-1+2j}{2}, \frac{2j+3}{2}; \frac{u}{2}\right) du\right].
\end{aligned}
$$

We will need the following formula (7.621.6) in Gradshteyn & Ryzhik (2007)

$$
\int_0^{\infty} t^{b-1} e^{-st} \Psi(a,c;t) = \frac{\Gamma(b)\Gamma(b-c+1)}{\Gamma(a+b-c+1)} {}_2F_1\left(b, b-c+1; a+b-c+1; 1-s\right) \tag{2.34}
$$

to get

$$
\begin{aligned}
N = & e^{-\lambda/2} \sum_{j=0}^{\infty} \frac{(\lambda/2)^j}{j!} 2^{-\frac{p-1+2j}{2}} \left\{\sum_{m=0}^{\infty} \frac{e^{-\frac{x_i}{2}} \left(\frac{x_i}{2}\right)^m}{m!} \frac{1}{\Gamma\left(\frac{1+2m}{m}\right) 2^{\frac{1+2m}{2}}} \right. \\
& \times \left[x_i \frac{\Gamma(j+m+1)\Gamma\left(m+\frac{1}{2}\right)}{\Gamma\left(\frac{p-3}{2}+m+j\right)} {}_2F_1\left(j+m+1, m+\frac{1}{2}; \frac{p-3}{2}+m+j; \frac{1}{2}\right)\right. \\
& \left.\left. - \frac{\Gamma\left(j+m+\frac{3}{2}\right)\Gamma(m+1)}{\Gamma\left(\frac{p-2}{2}+m+j\right)} {}_2F_1\left(j+m+\frac{3}{2}, m+1; \frac{p-2}{2}+m+j; \frac{1}{2}\right)\right]\right\}.
\end{aligned} \tag{2.35}
$$

The generalized Bayes rule under harmonic prior $g(\|\boldsymbol{\theta}\|) = \|\boldsymbol{\theta}\|^{2-p}$ is given by

$$\delta_g(\mathbf{x}) = \sum_{i=1}^{p} \left(1 - \frac{N}{D}\right) I(x_i = x_{(1)}), \tag{2.36}$$

where N is defined in (2.35) and D is defined in (2.33) for odd $p \geq 5$, and in (2.32) for even $p \geq 4$.

2.3.2 Generalized Bayes rules under the horseshoe prior

As in previous section, we want to construct the generalized Bayes rule (2.21). Here we consider horseshoe prior first introduced in Carvalho et al. (2010) for sparsity problems

$$\begin{aligned} X|\theta &\sim N(\theta, I_n), \\ \theta_i|\lambda_i, \tau &\sim N(0, \tau^2\lambda_i^2), \\ \lambda_i &\sim C^+(0, 1), \end{aligned} \tag{2.37}$$

where $C^+(0, a)$ is a standard half-Cauchy distribution on positive reals with scale parameter a.

Carvalho et al. (2010) compared the horseshoe prior to Cauchy, double-exponential, Strawderman-Berger, normal-exponential-gamma and normal-Jeffreys priors. The advantages of the horseshoe prior include tail robustness, that is, it shrinks observations close to zero much more than those far from zero. It also guarantees that the Bayes estimator for the sampling density converges efficiently to the truth.

We assume that $\tau = \frac{p_n}{p}$ is fixed (as in van der Pas et al. (2014)), where p_n is the number of non-zero parameters. In practice p_n might be unknown, and in this case we want to use empirically estimated τ like in van der Pas et al. (2014). Here we

will estimate $\tau = \max\left\{\hat{\tau}, \frac{1}{p}\right\}$ in the following way

$$\hat{\tau} = \frac{\#\left\{|x_i| \geq \sqrt{c_1\sigma^2 \log p}\right\}}{c_2 p}, \tag{2.38}$$

where c_1 and c_2 are positive constants. For the simulations we will use $c_1 = 2$ and $c_2 = 1$ as in van der Pas et al. (2014).

Then marginal density of θ_i under the horseshoe prior is given by

$$g(\theta_i) = \int_0^\infty \frac{1}{\sqrt{2\pi\lambda_i^2\tau^2}} \exp\left(-\frac{\theta_i^2}{2\lambda_i^2\tau^2}\right) \frac{2}{\pi(1+\lambda_i^2)}\, d\lambda_i. \tag{2.39}$$

By making the following transformations sequentially

$$u = \frac{1}{\lambda_i^2\tau^2},$$
$$z = \tau u + \frac{1}{\tau},$$

we get the following form for marginal distribution of θ_i

$$g(\theta_i) = \frac{1}{\tau\sqrt{2\pi^3}} \int_{1/\tau}^\infty \frac{1}{z} \exp\left(-\frac{\left(z - \frac{1}{\tau}\right)\theta_i^2}{2\tau}\right) dz.$$

We will use this form of generalized Bayes estimate

$$\delta_g(\mathbf{x}) = \sum_{i=1}^p \left[x_i + \frac{\int (\theta_i - x_i) p(\mathbf{x}|\boldsymbol{\theta}) g(\boldsymbol{\theta})\, d\boldsymbol{\theta}}{\int p(\mathbf{x}|\boldsymbol{\theta}) g(\boldsymbol{\theta})\, d\boldsymbol{\theta}}\right] I(x_i = x_{(1)}).$$

First, consider numerator separately

$$\begin{aligned}
&\int (\theta_i - x_i) p(\mathbf{x}|\boldsymbol{\theta}) g(\boldsymbol{\theta})\, d\boldsymbol{\theta} = \\
&= \int_\Theta (\theta_i - x_i) \int_{1/\tau}^\infty \left(\frac{1}{2\pi}\right)^{p/2} K^p \frac{1}{\tau^p} \frac{1}{\prod z_j} \exp\left\{-\frac{1}{2}\sum_{j=1}^p \frac{x_j^2 \frac{1}{\tau}(z_j - \frac{1}{\tau})}{\frac{1}{\tau}(z_j - \frac{1}{\tau}) + 1}\right\} d\boldsymbol{\theta} \\
&\qquad \times \exp\left\{-\frac{1}{2}\sum_{j=1}^p \left(\frac{1}{\tau}\left(z_j - \frac{1}{\tau}\right) + 1\right)\left(\theta_j - \frac{x_j}{\frac{1}{\tau}(z_j - \frac{1}{\tau}) + 1}\right)^2\right\} d\mathbf{z} \\
&= K^p \frac{1}{\tau^p} \prod_{\substack{j=1 \\ j\neq i}}^p \left\{\int_{1/\tau}^\infty \frac{1}{z} \frac{1}{\sqrt{\frac{1}{\tau}(z - \frac{1}{\tau}) + 1}} \exp\left(-\frac{1}{2}\frac{x_j^2 \frac{1}{\tau}(z - \frac{1}{\tau})}{\frac{1}{\tau}(z - \frac{1}{\tau}) + 1}\right) dz\right\} \\
&\qquad \times x_i \int_{1/\tau}^\infty \frac{1}{z} \frac{1}{\sqrt{\frac{1}{\tau}(z - \frac{1}{\tau}) + 1}} \left(-\frac{\frac{1}{\tau}(z - \frac{1}{\tau})}{\frac{1}{\tau}(z - \frac{1}{\tau}) + 1}\right) \exp\left(-\frac{1}{2}\frac{x_i^2 \frac{1}{\tau}(z - \frac{1}{\tau})}{\frac{1}{\tau}(z - \frac{1}{\tau}) + 1}\right) dz,
\end{aligned}$$

where $K = \frac{1}{\sqrt{2\pi^3}}$.

Now consider the denominator

$$\int p(\mathbf{x}|\boldsymbol{\theta})g(\boldsymbol{\theta})\,d\boldsymbol{\theta} = \int_\Theta \int_{1/\tau}^\infty \left(\frac{1}{2\pi}\right)^{p/2} \left(\frac{1}{2\pi^3}\right)^{p/2} \exp\left(-\frac{1}{2}\sum_{j=1}^p \frac{x_j^2 \frac{1}{\tau}(z_j - \frac{1}{\tau})}{\frac{1}{\tau}(z_j - \frac{1}{\tau}) + 1}\right)$$
$$\times \exp\left\{-\frac{1}{2}\sum_{j=1}^p \left(\frac{1}{\tau}\left(z_j \frac{1}{\tau}\right) + 1\right)\left(\theta_j - \frac{x_j}{\frac{1}{\tau}(z_j - \frac{1}{\tau}) + 1}\right)^2\right\} d\mathbf{z}$$
$$= K^p \frac{1}{\tau^p} \prod_{j=1}^p \left\{\int_{1/\tau}^\infty \frac{1}{z} \frac{1}{\sqrt{\frac{1}{\tau}(z - \frac{1}{\tau}) + 1}} \exp\left(-\frac{1}{2}\frac{x_j^2 \frac{1}{\tau}(z - \frac{1}{\tau})}{\frac{1}{\tau}(z - \frac{1}{\tau}) + 1}\right)\, dz\right\}.$$

Thus, the generalized Bayes rule under the horseshoe prior is given by

$$\delta_g(\mathbf{x}) = \sum_{i=1}^p \left[1 - \frac{\int_{1/\tau}^\infty \frac{1}{z} \frac{1}{\sqrt{\frac{1}{\tau}(z-\frac{1}{\tau})+1}} \left(-\frac{\frac{1}{\tau}(z-\frac{1}{\tau})}{\frac{1}{\tau}(z-\frac{1}{\tau})+1}\right) \exp\left(-\frac{1}{2}\frac{x_i^2 \frac{1}{\tau}(z-\frac{1}{\tau})}{\frac{1}{\tau}(z-\frac{1}{\tau})+1}\right)\, dz}{\int_{1/\tau}^\infty \frac{1}{z} \frac{1}{\sqrt{\frac{1}{\tau}(z-\frac{1}{\tau})+1}} \exp\left(-\frac{1}{2}\frac{x_i^2 \frac{1}{\tau}(z-\frac{1}{\tau})}{\frac{1}{\tau}(z-\frac{1}{\tau})+1}\right)\, dz}\right] x_i I(x_i = x_{(1)}).$$

To simplify the form of δ_g, make the transformation $\frac{\frac{z\tau-1}{\tau^2}}{\frac{z\tau-1}{\tau^2}+1} = t$, i.e. $z = \frac{1-t(1-\tau^2)}{\tau(1-t)}$,

$$\delta_g(\mathbf{x}) = \sum_{i=1}^p \left[1 - \frac{\int_0^1 \frac{t}{\sqrt{1-t(1-t(1-\tau^2))}} \exp\left(-\frac{x_i^2 t}{2}\right)\, dt}{\int_0^1 \frac{1}{\sqrt{1-t(1-t(1-\tau^2))}} \exp\left(-\frac{x_i^2 t}{2}\right)\, dt}\right] x_i I(x_i = x_{(1)})$$
$$= \sum_{i=1}^p \left[1 - \frac{2\Phi_1\left(\frac{1}{2}, 1, \frac{5}{2}; \frac{x_i^2}{2}, 1 - \frac{1}{\tau^2}\right)}{3\Phi_1\left(\frac{1}{2}, 1, \frac{3}{2}; \frac{x_i^2}{2}, 1 - \frac{1}{\tau^2}\right)}\right] x_i I(x_i = x_{(1)}). \tag{2.40}$$

2.3.3 Competing estimators

In this section we will compare the performance of the competing estimators proposed in the literature and the Bayes estimators suggested above.

Estimator of Cohen & Sackrowitz (1982)

The estimator mentioned earlier in the introduction, T_2, for estimating $\theta_{(1)}$, suggested in Cohen & Sackrowitz (1982), has a bias smaller than that of $X_{(1)}$. The

estimator T_2 is defined in the following way

$$T_2(\mathbf{X}) = \sum_{i=1}^{p} C_{i,p} X_{(i)},$$

where

$$C_{i,p} = \begin{cases} \frac{\beta_i - 1}{\beta_1 \beta_2 \ldots \beta_i} & \text{for } i = 1, 2, \ldots, p-1, \\ \frac{1}{\beta_1 \beta_2 \ldots \beta_{p-1}} & \text{for } i = p, \end{cases}$$

with

$$\beta_{p-1} = \hat{r}_{p-1} + 2,$$
$$\beta_i = \hat{r}_i + 2 - \frac{1}{\beta_{i+1}} \text{ for } i = 1, 2, \ldots, p-2,$$

where $\hat{r}_i = 2\left(X_{(i)} - X_{(i+1)}\right)^2$, for $i = 1, 2, \ldots, p-1$ and $C_{i,p}$ are such that $C_{1,p} \geq C_{2,p} \geq \cdots \geq C_{p-1,p}$.

The estimator of Reid & Tibshirani (2014)

Reid & Tibshirani (2014) implicitly make an assumption of sparsity, with many effect sizes $\theta_i = 0$. They worked on the problem of estimating means conditional on the selected k largest (absolute) order statistics. Currently there are two ways of dealing with selection bias and inference after selection: empirical Bayes approach (e.g. Efron (2011)) and classical resampling approach (e.g. Simon & Simon (2013)). Reid & Tibshirani (2014) proposed method is based on theory developed in Lee et al. (2014) where they consider doing post-selection inference with the Lasso. They assume

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon} \text{ with } \mathbf{X} \in R^{n \times p} \text{ and } \boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}).$$

In Reid & Tibshirani (2014) they consider the orthogonal case $\mathbf{X} = \mathbf{I}$. They define $E = \{r(k) : k = 1, \ldots, K\}$, where $r(k)$ defines the permutation such that $\left|Y_{r(k)}\right| = |Y|_{(k)}$, and E is set of non-zero (or selected) indexes. Define

$z_{E,j} = \text{sign}\left(\hat{\theta}_{\lambda,j}\right)$, where $\hat{\theta}_{\lambda,j}$ is the solution to the Lasso problem (with some regularization parameter λ). They claim doing inference after selection is equivalent to conditioning on $\{E, z_E, G, z_G\} = \{A_K \mathbf{y} \leq b_K\}$, where $G = \{r(K+1)\}$, that is the selection event is equivalent putting some affine constraints on $\mathbf{y}$. For definitions of A_K and b_K see (5) in Reid & Tibshirani (2014). Under these assumptions they derive the distribution of $y_i, i \in E$ as a truncated Gaussian distribution on $(-\infty, -|y|_{(K+1)}] \cup [|y|_{(K+1)}, \infty)$, and they arrive at the conclusion that point estimate $\hat{\theta}_{r(k)}$ should satisfy

$$y_{r(k)} = \hat{\theta}_{r(k)} + \frac{\phi\left(|y|_{K+1} - \hat{\theta}_{r(k)}\right) - \phi\left(|y|_{K+1} + \hat{\theta}_{r(k)}\right)}{\Phi\left(-|y|_{K+1} - \hat{\theta}_{r(k)}\right) + 1 - \Phi\left(|y|_{r(k)} - \hat{\theta}_{r(k)}\right)}. \tag{2.41}$$

2.4 Numerical studies

In this section we will provide a numerical comparison of the suggested admissible estimator above (2.40), which is a generalized Bayes estimator under the horseshoe prior (generalized Bayes estimator under the harmonic prior performs very similarly to the naive estimator, and hence is omitted), standard estimators such as naive estimator and James-Stein estimator (JS), and estimators suggested specifically for the selection problem Cohen & Sackrowitz (1982) (CS) and Reid & Tibshirani (2014). Since estimator in Cohen & Sackrowitz (1982) was only given for the case of selecting one population we divide this section in two parts. In the first part we will carry out simulations for the case of selecting one population with the largest sample mean. In the second part we provide numerical results for general k. All the simulations to follow are performed in the view of sparsity of $\boldsymbol{\theta}$.

2.4.1 Case $k = 1$

The estimator (2.41) of Reid & Tibshirani (2014) is the only one out of the considered estimators that is defined for absolute value selection, and hence is not included in comparison for the case $k = 1$.

The following table provides simulations of mean squared error for different levels of sparsity and signal strength.

Table 2.1: MSE over 1000 simulations, $p = 100$.

$\boldsymbol{\theta}$	Naive	JS	CS	Horseshoe
$\theta_1 = \cdots = \theta_p = 0$	7.68	0.16	3.92	0.15
$\theta_1 = \cdots = \theta_p = 5$	6.37	5.03	3.8	5.1
$\theta_1 = 0,\ \theta_2 = 0.1,\ \theta_3 = 0.2, \ldots,\ \theta_p = 0.1p$	3.29	2.32	1.96	2.73
$\theta_1 = 0,\ \theta_2 = 1,\ \theta_3 = 2, \ldots,\ \theta_p = p$	1.23	1.2	1.04	1.11
$\theta_1 = \cdots = \theta_{0.5p} = 0$ and $\theta_{0.5p+1} = \cdots = \theta_p = 1$	6.06	0.3	3.58	2.6
$\theta_1 = \cdots = \theta_{0.5p} = 0$ and $\theta_{0.5p+1} = \cdots = \theta_p = 5$	5.24	3.14	3.12	4.03
$\theta_1 = \cdots = \theta_{0.5p} = 0$ and $\theta_{0.5p+1} = \cdots = \theta_p = 10$	5.27	5.09	3.14	5.04
$\theta_1 = \cdots = \theta_{0.75p} = 0$ and $\theta_{0.75p+1} = \cdots = \theta_p = 1$	6.16	0.35	3.54	1.85
$\theta_1 = \cdots = \theta_{0.75p} = 0$ and $\theta_{0.75p+1} = \cdots = \theta_p = 5$	4.11	1.26	2.44	3.02
$\theta_1 = \cdots = \theta_{0.75p} = 0$ and $\theta_{0.75p+1} = \cdots = \theta_p = 10$	4.12	3.81	2.44	3.92
$\theta_1 = \cdots = \theta_{0.9p} = 0$ and $\theta_{0.9p+1} = \cdots = \theta_p = 1$	6.66	0.39	3.7	1.18
$\theta_1 = \cdots = \theta_{0.9p} = 0$ and $\theta_{0.9p+1} = \cdots = \theta_p = 5$	2.78	0.3	1.64	1.88
$\theta_1 = \cdots = \theta_{0.9p} = 0$ and $\theta_{0.9p+1} = \cdots = \theta_p = 10$	2.9	2.34	1.74	2.73

As expected, the naive estimate $X_{(1)}$ performs the worst. The only case when it performs comparably to the other estimates is the case when $\theta_1 = 0, \theta_2 = 1, \ldots, \theta_p = p$. Since we are only considering selecting one parameter, naive estimator could pick it up reasonably well, since those values of θ are well separated. For very sparse configurations James-Stein estimator, which is not specifically designed for the selection problem, outperforms both CS and horseshoe estimates, which are constructed under selection. CS outperforms horseshoe for less sparse scenarios when signals are well separated, while horseshoe performs much better under sparsity.

2.4.2 General k

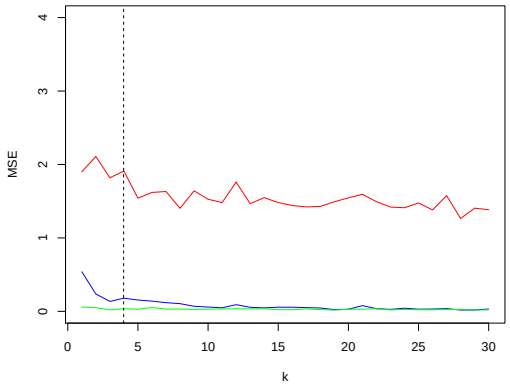
For simulations we use the same set up as in Reid & Tibshirani (2014). Samples of size p are generated from $N(\theta, I)$ distribution. Sparsity of θ is controlled by parameter λ, where $\lambda \in [0, 1]$, so that $\lceil p^{\lambda} \rceil$ entries are non-zero. Thus, lower λ corresponds to more sparsity. Non-zero entries were generated from $N(\nu, 1)$ distribution. We considered the following values of the parameters:

- Sample size $p = 100, 200, 500, 1000$.
- Sparsity parameter $\lambda = 0.2, 0.3, 0.45, 0.7, 1$.
- Signal strength parameter $\nu = 0, 1, \ldots, 5, 10$.

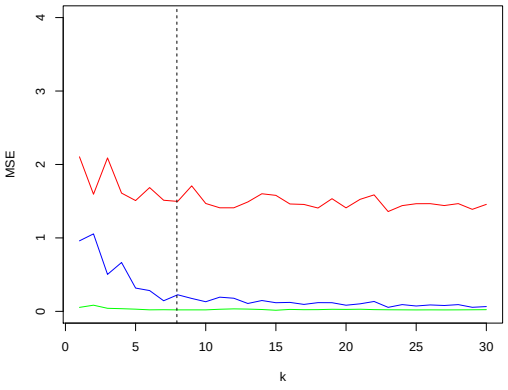
Since estimator of Reid & Tibshirani (2014) is defined for the absolute value selection, in the following plots all procedures were adjusted to selection corresponding to the largest absolute value sample means. Simulations for the selection corresponding to the largest sample mean are provided in the Section 2.5.

In Figure 2.1 the signal is very small and close to zero. For that reason for sparse situations James-Stein estimator mostly dominates other estimators. The naive estimate perform very poorly for sparse scenarios, its curves often running off the top of the plot. For levels of sparsity up to $\lambda = 0.45$, the horseshoe estimator uniformly dominates Reid's estimator. For less sparse situation with $\lambda = 0.7$, the horseshoe procedure performs worse, but eventually dominates Reid. For no sparsity scenario $\lambda = 1$, which is not of interest, Reid's estimator outperforms other estimators for very small signals.

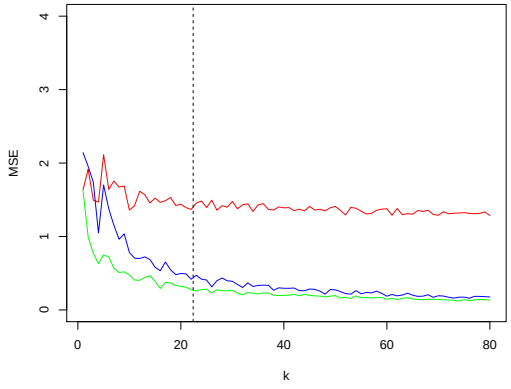
In Figure 2.2 the general behavior seems to be that horseshoe estimator dominates Reid's estimator when $k < \lceil p^{\lambda} \rceil$. Performance of Reid & Tibshirani (2014) estimator improves as k approaches and moves beyond the true number of non-zero



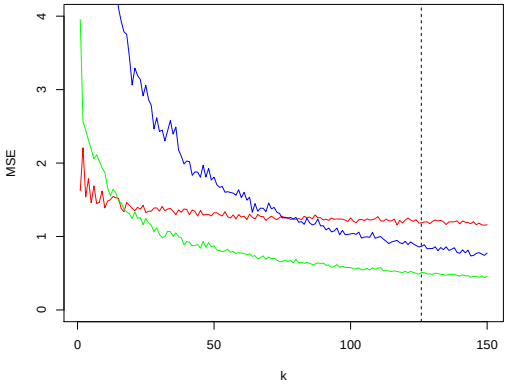
(a)



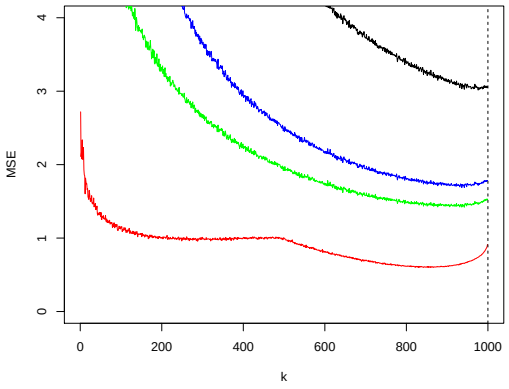
(b)



(c)



(d)



(e)

Figure 2.1: Mean squared error as a function of the number of selected populations k. Sparsity varies over panels. (a) $\lambda = 0.2$, (b) $\lambda = 0.3$, (c) $\lambda = 0.45$, (d) $\lambda = 0.7$, (e) $\lambda = 1$. Vertical dotted lines at p^{λ} - the true number of non-zero signals. Sample size $p = 1000$, signal size $\nu = 0$. MSE for naive estimates is above 8, and is omitted. Black lines correspond to the naive estimates, blue - generalized Bayes estimates under horseshoe prior, red - Reid's estimates, green - James-Stein estimates.

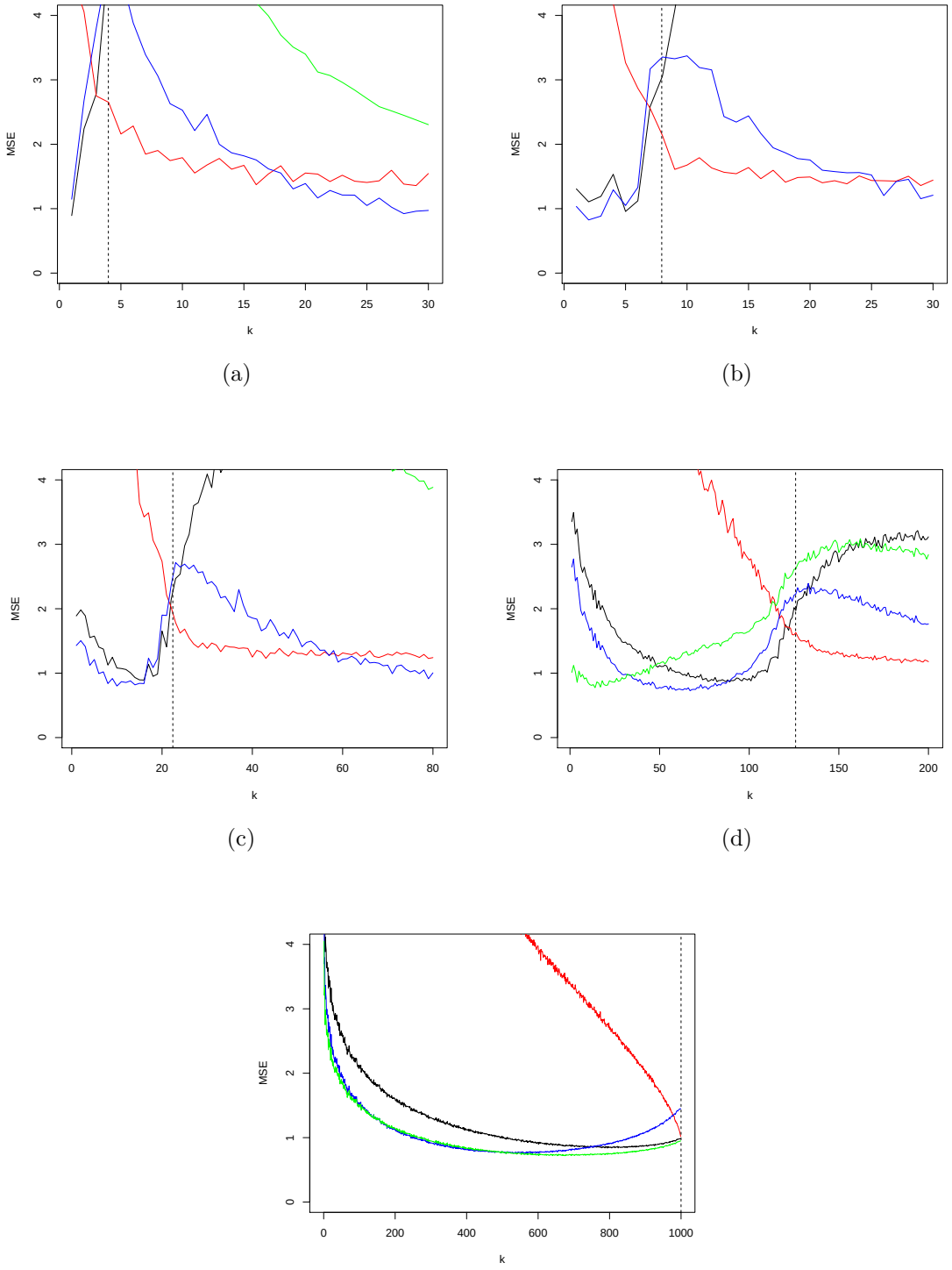


Figure 2.2: Mean squared error as a function of the number of selected populations k. Sparsity varies over panels. (a) $\lambda = 0.2$, (b) $\lambda = 0.3$, (c) $\lambda = 0.45$, (d) $\lambda = 0.7$, (e) $\lambda = 1$. Vertical dotted lines at p^λ - the true number of non-zero signals. Sample size $p = 1000$, signal size $\nu = 5$. MSE for naive estimates is above 8, and is omitted. Black lines correspond to the naive estimates, blue - generalized Bayes estimates under horseshoe prior, red - Reid's estimates, green - James-Stein estimates.

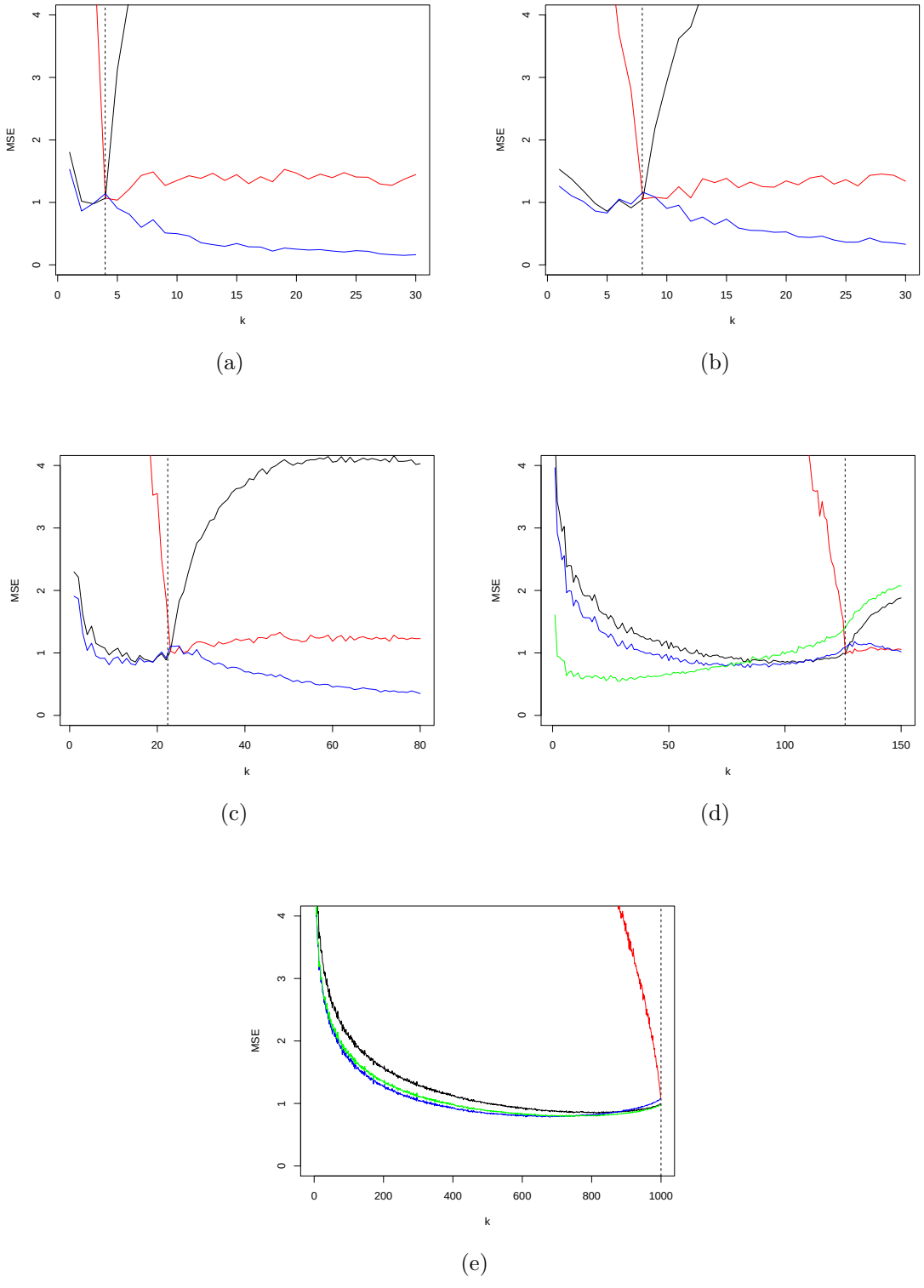
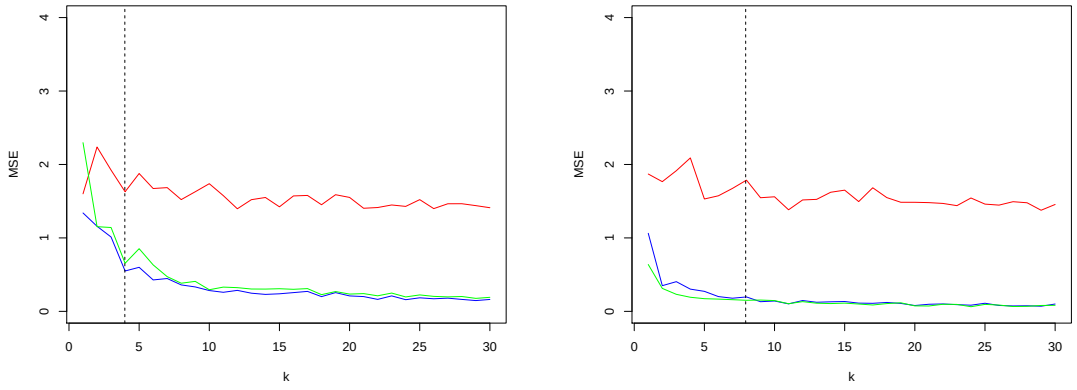


Figure 2.3: Mean squared error as a function of the number of selected $\lambda = 0.2$, (b) $\lambda = 0.3$, (c) $\lambda = 0.45$, (d) $\lambda = 0.7$, (e) $\lambda = 1$. Vertical dotted lines at p^{λ} - the true number of non-zero signals. Sample size $p = 1000$, signal size $\nu = 10$. MSE for naive estimates is above 8, and is omitted. Black lines correspond to the naive estimates, blue - generalized Bayes estimates under horseshoe prior, red - Reid's estimates, green - James-Stein estimates.

signals, but eventually horseshoe estimator will dominate it. From Figure 2.3 we can see that for very strong non-zero signals $\nu = 10$ horseshoe estimator dominates Reid's estimator for all levels of sparsity.

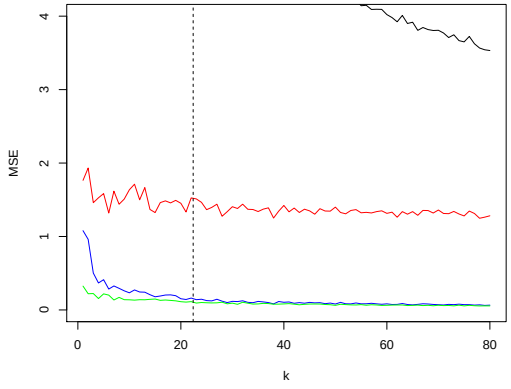
2.5 Additional plots: non-absolute value selection

Since generalized Bayes rule (2.21) was defined for selection corresponding to the largest mean, we provide simulations for this type of selection below.

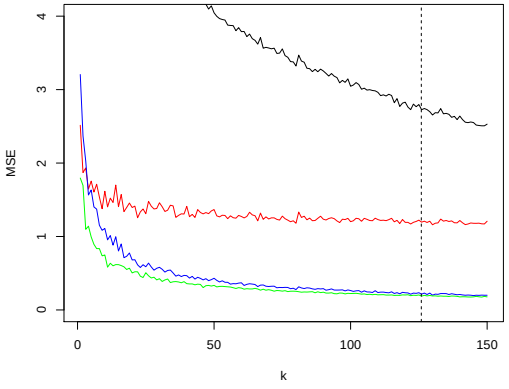


(a)

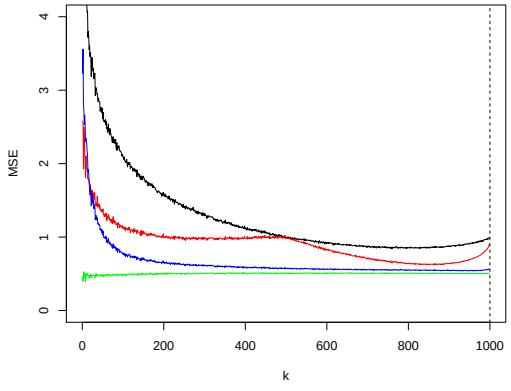
(b)



(c)



(d)



(e)

Figure 2.4: Mean squared error as a function of the number of selected populations k. Sparsity varies over panels. (a) $\lambda = 0.2$, (b) $\lambda = 0.3$, (c) $\lambda = 0.45$, (d) $\lambda = 0.7$, (e) $\lambda = 1$. Vertical dotted lines at p^{λ} - the true number of non-zero signals. Sample size $p = 1000$, signal size $\nu = 0$. MSE for naive estimates is above 8, and is omitted. Black lines correspond to the naive estimates, blue - generalized Bayes estimates under horseshoe prior, red - Reid's estimates, green - James-Stein estimates.

CHAPTER 3

COVERAGE PROBABILITY

In this section, as before, consider p-dimensional random vector $\mathbf{X} \sim N_p(\boldsymbol{\theta}, \mathbf{I})$.

The naive set (interval) for estimating $\theta_{(1)} = \sum_{i=1}^{p} X_i I(X_i = X_{(1)})$ is given by

$$C_0 = \left\{\theta_{(1)} : (X_{(1)} - \theta_{(1)})^2 \leq h^2\right\}, \tag{3.1}$$

where $h = z_{1-\alpha/2}$. This set can be rewritten in the way that selection becomes more apparent

$$C_0 = \left\{\theta_{(1)} : \sum_{i=1}^{p} (X_i - \theta_i)^2 I(X_i = X_{(1)}) \leq h^2\right\}. \tag{3.2}$$

For the case when we are interested in estimating k selected parameters $\boldsymbol{\theta}_s = \left(\theta_{(1)}, \ldots, \theta_{(k)}\right)$ simultaneously, the naive confidence set is

$$C_0 = \left\{(\theta_{(1)}, \ldots, \theta_{(k)}) : \sum_{i=1}^{p} \sum_{j=1}^{k} (X_i - \theta_i)^2 I(X_i = X_{(j)}) \leq h^2\right\}, \tag{3.3}$$

where $h^2 = \chi^2_{1-\alpha,k}$. Simultaneous confidence sets, in addition to controlling the levels of the individual confidence statements that make up the confidence set, need to control the overall level $1 - \alpha$.

Improved confidence intervals for $\theta_{(1)}$ were proposed in Venter (1988), Qiu & Hwang (2007), Efron (2011), Fuentes et al. (2014), and Reid & Tibshirani (2014). Here we concentrate on the simultaneous confidence sets. In particular, in this section we will evaluate the coverage probability for the naive sets and will use to obtain improved sets with guaranteed coverage probability $1 - \alpha$.

3.1 Exploring the behavior of the usual confidence set

Before we give the theoretical results about the usual confidence set for estimating the normal mean in selection context, we will explore behavior of the naive set (3.3)

and demonstrate its limitations.

3.1.1 Selecting one population, $k = 1$

In the Figure 3.1 the coverage probability of the usual confidence set is plotted for different values of p, the number of populations considered. In the case for $p = 2$ the set maintains $1 - \alpha$ coverage probability for all $\boldsymbol{\theta}$ (in Section 3.2 we provide theoretical calculations for the coverage probability, which turns out for this case coverage probability is exactly $1 - \alpha$). As we increase the number of populations p the coverage probability goes to zero very fast for small $\|\boldsymbol{\theta}\|$. For larger $\|\boldsymbol{\theta}\|$ the coverage probability is maintained at the nominal level for all p. The reason is that as we increase $\|\boldsymbol{\theta}\|$, the individual components of $\boldsymbol{\theta}$ get further away from each other, and it is easier to distinguish between them to find selected θ. Thus usual confidence set performs satisfactorily for all of the components of $\boldsymbol{\theta}$ only for $p = 2$.

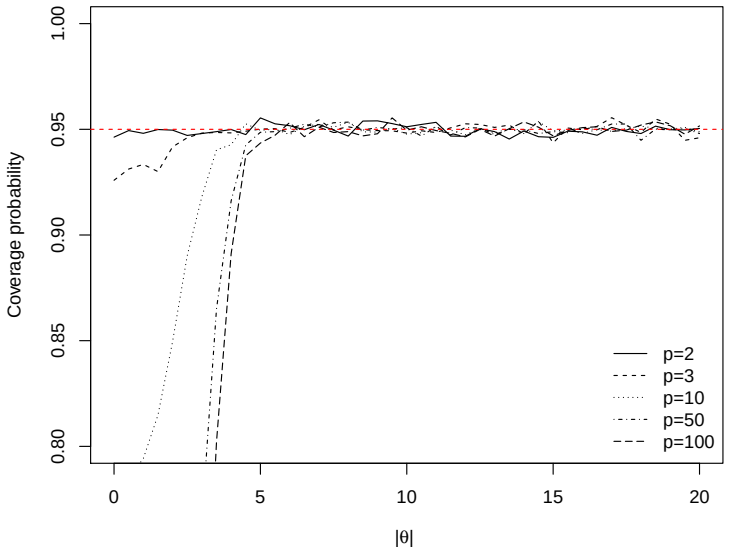
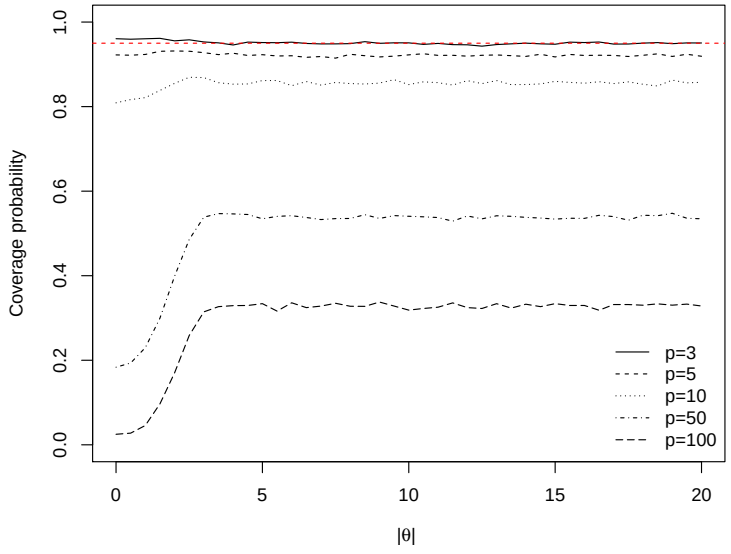


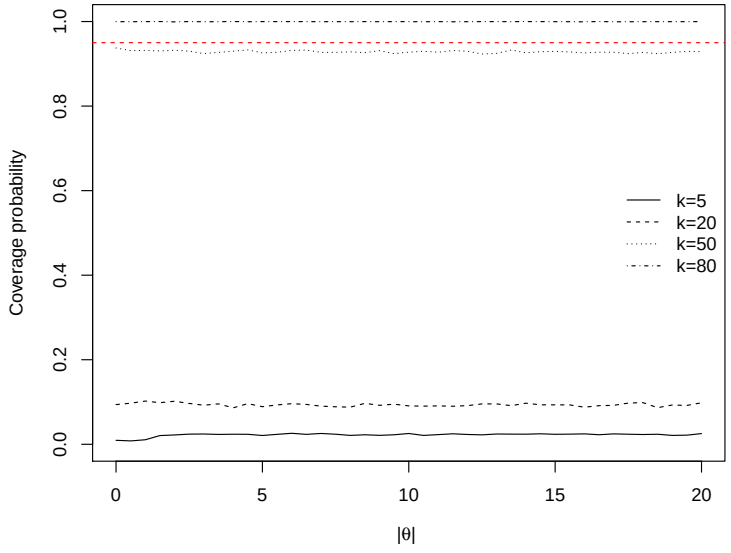
Figure 3.1: Coverage probability of the naive set (3.1) for $k = 1$.

3.1.2 Selecting several populations, general k

For the general case, $k > 1$, the coverage probability for the usual confidence set is more complex and is provided below. In the Figure 3.2 (a) we keep k being small, $k = 2$, and increase p, number of considered populations. Except for the case $p = 3$, the coverage probability is always less than $1 - \alpha$. In the Figure 3.2 (b) we set $p = 100$, and increase k, number of selected populations. As k increases, the coverage probability goes from zero, to almost perfect coverage when k is comparable with p. Unfortunately, this is not the case we are interested in. We are interested in selecting a few populations out of many which corresponds to small k compared to p. In all cases we see that the coverage probability is low for the high dimensional parameter spaces when the number of components to be selected is small. This is precisely the setting in estimation of sparse high dimensional parameter sets.



(a)



(b)

Figure 3.2: Coverage probability of the naive set (3.3) for (a) $k = 2$, (b) $p = 100$.

3.2 Exact coverage probability of C_0 for $k = 1$ and $p = 2$

Here we calculate the exact coverage probability for C_0 for the case $k = 1$ and $p = 2$.

$$P\left(\left|\theta_{(1)} - X_{(1)}\right| \le h\right) = \sum_{i=1}^{2} P\left(|\theta_i - X_i| \le h, X_i = X_{(1)}\right)$$

$$= P\left(|\theta_i - X_i| \le h, X_1 \ge X_2\right) + P\left(|\theta_i - X_i| \le h, X_2 \ge X_1\right)$$

$$= \int_{-h}^{h} \Phi\left(z - \Delta_{21}\right) \phi(z)\, dz + \int_{-h}^{h} \Phi\left(z - \Delta_{12}\right) \phi(z)\, dz$$

$$= \left[\left\{T\left(z, -\frac{\Delta_{21}}{z\sqrt{2}}\right) + T\left(-\frac{\Delta_{21}}{\sqrt{2}}, \frac{z\sqrt{2}}{\Delta_{21}}\right) - T\left(z, \frac{-\Delta_{21} + z}{z}\right) - T\left(-\frac{\Delta_{21}}{\sqrt{2}}, \frac{-\Delta_{21} + 2z}{-\Delta_{21}}\right)\right.\right.$$
$$\left. + \Phi(z)\Phi\left(-\frac{\Delta_{21}}{\sqrt{2}}\right)\right\} + \left\{T\left(z, -\frac{\Delta_{12}}{z\sqrt{2}}\right) + T\left(-\frac{\Delta_{12}}{\sqrt{2}}, -\frac{z\sqrt{2}}{\Delta_{12}}\right) - T\left(z, \frac{-\Delta_{12} + z}{z}\right)\right.$$
$$\left.\left. - T\left(-\frac{\Delta_{12}}{\sqrt{2}}, \frac{-\Delta_{12} + 2z}{-\Delta_{12}}\right) + \Phi(z)\Phi\left(-\frac{\Delta_{12}}{\sqrt{2}}\right)\right\}\right]\Bigg|_{-h}^{h},$$

where $Z = X_i - \theta_i$ and $\Delta_{ji} = \theta_j - \theta_i$.

Here we used the following facts (see e.g. Owen (1980)) that

$$T(h, a) = \int_0^a \frac{\phi(h)\phi(hx)}{1 + x^2}\, dx \text{ for } -\infty < h < \infty, \text{ and } 0 < a < \infty,$$

$$T(-h, a) = T(h, a),$$

$$T(h, -a) = -T(h, a).$$

By regrouping and simplifying the terms in the expression above, we get

$$P\left(\left|\theta_{(1)} - X_{(1)}\right| \le h\right) =$$
$$= \left[T\left(z, \frac{\Delta_{12}}{z\sqrt{2}}\right) + T\left(\frac{\Delta_{12}}{\sqrt{2}}, \frac{z\sqrt{2}}{\Delta_{12}}\right) - T\left(z, \frac{\Delta_{12} + z}{z}\right) - T\left(\frac{\Delta_{12}}{\sqrt{2}}, \frac{\Delta_{12} + 2z}{\Delta_{12}}\right)\right.$$
$$+ \Phi\left(z\right)\Phi\left(\frac{\Delta_{12}}{\sqrt{2}}\right) - T\left(z, \frac{\Delta_{12}}{z\sqrt{2}}\right) - T\left(\frac{\Delta_{12}}{\sqrt{2}}, \frac{z\sqrt{2}}{\Delta_{12}}\right) + T\left(z, \frac{\Delta_{12} - z}{z}\right)$$
$$\left. - T\left(\frac{\Delta_{12}}{\sqrt{2}}, \frac{\Delta_{12} - 2z}{\Delta_{12}}\right) + \Phi\left(z\right)\Phi\left(-\frac{\Delta_{12}}{\sqrt{2}}\right)\right]\Bigg|_{-h}^{h}$$

$$
\begin{aligned}
&= \left[T\left(z, \frac{\Delta_{12}-z}{z}\right) - T\left(z, \frac{\Delta_{12}+z}{z}\right) - T\left(\frac{\Delta_{12}}{\sqrt{2}}, \frac{\Delta_{12}+2z}{\Delta_{12}}\right) \right. \\
&\qquad \left. + \; T\left(\frac{\Delta_{12}}{\sqrt{2}}, \frac{\Delta_{12}-2z}{\Delta_{12}}\right) + \Phi(z) \right] \Bigg|_{-h}^{h} \\
&= T\left(h, \frac{\Delta_{12}-h}{h}\right) - T\left(h, \frac{\Delta_{12}+h}{h}\right) - T\left(\frac{\Delta_{12}}{\sqrt{2}}, \frac{\Delta_{12}+2h}{\Delta_{12}}\right) - T\left(\frac{\Delta_{12}}{\sqrt{2}}, \frac{\Delta_{12}-2h}{\Delta_{12}}\right) \\
&\qquad + \Phi(h) - T\left(-h, \frac{\Delta_{12}+h}{-h}\right) + T\left(-h, \frac{\Delta_{12}-h}{-h}\right) + T\left(\frac{\Delta_{12}}{\sqrt{2}}, \frac{\Delta_{12}-2h}{\Delta_{12}}\right) \\
&\qquad + T\left(\frac{\Delta_{12}}{\sqrt{2}}, \frac{\Delta_{12}+2c}{\Delta_{12}}\right) - \Phi(-h) \\
&= \Phi(h) - \Phi(-h) \\
&= 1 - \alpha.
\end{aligned}
$$

Thus, for selecting one population out of two, the naive set C_0 maintains coverage probability $1-\alpha$. For the case $p > 2$ it is difficult to get an analytical expression for the integrals involved in the expression for coverage probability. But numerically, as in the Figure 3.1, we can see that C_0 fails to maintain nominal coverage probability, particularly for small values of $\|\boldsymbol{\theta}\|$.

3.2.1 Coverage probability bound for C_0 for $k > 1$

The idea behind this section is to get a lower bound on the coverage probability of the naive set. By equating this lower bound to $1-\alpha$, we can find the radius of the set h^* that guarantees coverage $1-\alpha$.

For the case $k = 1$ Fuentes et al. (2014) proved that

$$
\min_{\theta} P\left(|X_{(1)} - \theta_{(1)}| \leq h)\right) = p \int_{-h}^{h} \Phi^{p-1}(z)\phi(z)\, dz = \Phi^p(h) - \Phi^p(-h).
$$

The set which dominates the naive set in this case is given by $C = \left\{\theta_{(1)} : |X_{(1)} - \theta_{(1)}| \leq h^*\right\}$, where h^* satisfies

$$
\Phi^p(h^*) - \Phi^p(-h^*) = 1 - \alpha. \tag{3.4}
$$

We start by looking at the case $k = 2$, that is we are interested in constructing confidence interval for $\boldsymbol{\theta}_s = (\theta_{(1)}, \theta_{(2)})$. In this case, as in Fuentes et al. (2014) we do not need to know the specific configuration of the vector $(X_{(1)}, \ldots, X_{(p)})$, we only need to consider $\binom{p}{k}$ different ways to select k out of p populations without considering the order. Suppose $X_{j_1}, \ldots, X_{j_k}$ correspond to the selected k variables, and $X_{j_{k+1}}, \ldots, X_{j_p}$ to non-selected $p-k$ variables. We can write coverage probability as

$$P(\|\boldsymbol{\theta}_s - \mathbf{X}_s\|^2 \leq h^2) = P\left(\sum_{i=1}^{2}(\theta_{(i)} - X_{(i)})^2 \leq h^2\right)$$
$$= \sum_{m=1}^{\binom{p}{k}} P\left(\sum_{i=1}^{2}(\theta_{j_i} - X_{j_i})^2 \leq h^2, \min(X_{j_1}, X_{j_2}) \geq \max(X_{j_3}, \ldots, X_{j_p})\right).$$

For example, for the case $(j_1, \ldots, j_k) = (1, \ldots, k)$, the corresponding term in the sum is

$$P\left(\sum_{i=1}^{2}(\theta_i - X_i)^2 \leq h^2, \min(X_1, X_2) \geq \max(X_3, \ldots, X_p)\right) =$$
$$= P\left(\sum_{i=1}^{2}(\theta_i - X_i)^2 \leq h^2, X_1 \geq X_3, \ldots, X_1 \geq X_p, X_2 \geq X_3, \ldots, X_2 \geq X_p\right) \tag{3.5}$$
$$= \int_{z_1^2+\cdots+z_k^2 \leq h^2} \cdots \int \prod_{j=k}^{p} \Phi\left(\min\{z_1 + \theta_1, \ldots, z_k + \theta_k\} - \theta_j\right) \phi(z_1) \ldots \phi(z_k)\, dz_1 \ldots\, dz_k \tag{3.6}$$

Define the centered variables $Z_j = X_j - \theta_j$ for $j = 1, \ldots, p$, then it is clear that $Z_1, \ldots, Z_p$ are iid $N(0, 1)$. It follows that

$$X_1 \geq X_j \Leftrightarrow X_1 - \theta_1 \geq X_j - \theta_j + \theta_j - \theta_1$$
$$Z_1 \geq Z_j + \Delta_{j1}$$
$$Z_1 - Z_j \geq \Delta_{j1},$$

where $\Delta_{j1} = \theta_j - \theta_1$ for $j = 1, \ldots, p$.

Then the full coverage probability is given by

$$P(\|\boldsymbol{\theta}_s - \mathbf{X}_s\|^2 \le h^2) = \sum_{j=1}^{\binom{p}{k}} \int \cdots \int_{\substack{\sum \\ q\in I_j}\ z_q^2 \le h^2} \prod_{m\in I_j^C} \Phi\left(\min_{l\in I_j}\{z_l + \theta_l\} - \theta_m\right) \prod_{l\in I_j} \phi(z_l)\, dz_l, \tag{3.7}$$

where $I_j = \{j_1, \ldots, j_k\}$ are selected indices and $I_j^C = \{j_{k+1}, \ldots, j_p\}$ are non-selected indices.

We want to get a sharp lower bound on the coverage probability (3.7). We provide two different ways to estimate this lower bound. The first way is a very simple approximation that can be considered as the first order approximation. The second approximation is more involved, but in the end it produces confidence sets with good coverage probability and reasonable size.

Approximation using integration over inscribed hypercube

We will first approximate integrating over a k-sphere in (3.7) with integrating over a hyper-cube inscribed into this k-sphere.The inscribed hyper-cube will have a side of the size $\frac{2h}{\sqrt{k}}$.

$$\begin{aligned} P\left(\|\mathbf{X}_s - \boldsymbol{\theta}_s\|^2 \le h^2\right) &\ge \sum_{j=1}^{\binom{p}{k}} \int_{-h/\sqrt{k}}^{h/\sqrt{k}} \cdots \int_{-h/\sqrt{k}}^{h/\sqrt{k}} \prod_{m\in I_j^c} \Phi\left(\min_{l\in I_j}\{z_l - \theta_l\} - \theta_m\right) \prod_{l\in I_j} \phi(z_l)\, dz_l \\ &= \left(\Phi\left(\frac{h}{\sqrt{k}}\right) - \Phi\left(-\frac{h}{\sqrt{k}}\right)\right)^{k-1} \left[\Phi^{p-k+1}\left(\frac{h}{\sqrt{k}}\right) - \Phi^{p-k+1}\left(-\frac{h}{\sqrt{k}}\right)\right]. \end{aligned} \tag{3.8}$$

To guarantee the coverage probability to be $1-\alpha$, we need to have the above approximation be greater than $1-\alpha$. Using this approximation we have the following result.

Proposition 1. *For a given p and k the usual confidence set $C_0 = \left\{\boldsymbol{\theta}_s : \sum_{i=1}^p \sum_{j=1}^k (X_i - \theta_i)^2 I(X_i = X_{(j)}) \le h^2\right\}$ has coverage probability of $1-\alpha$ for*

h the solution of

$$\left(\Phi\left(\frac{h}{\sqrt{k}}\right)-\Phi\left(-\frac{h}{\sqrt{k}}\right)\right)^{k-1}\left[\Phi^{p-k+1}\left(\frac{h}{\sqrt{k}}\right)-\Phi^{p-k+1}\left(-\frac{h}{\sqrt{k}}\right)\right]=1-\alpha. \quad (3.9)$$

Approximation using quadrature rules

The second approximation uses some nice results from approximation theory by Mustard (1964).

Due to the symmetry of the problem, assume $\theta_1 \leq \theta_2 \leq \cdots \leq \theta_p$. For now for simplicity consider only the term in the sum in (3.7) that corresponds to $I_j = \{1,\ldots,k\}$ and $I_j^C = \{p-k+1,\ldots,p\}$. In the following we will use similar approach to Mustard (1964) approximate this integral over a k-sphere. Define $I = \int_D f(\mathbf{z})\,d\mathbf{z}$.The idea is to find the following approximation using quadrature rules

$$I = \underset{\sum_{q\in\{1,\ldots,k\}} z_q^2 \leq h^2}{\int\cdots\int} \prod_{m\in\{p-k+1,\ldots,p\}} \Phi\left(\min_{l\in\{1,\ldots,k\}}\{z_l+\theta_l\}-\theta_m\right) \prod_{l\in\{1,\ldots,k\}} \phi(z_l)\,dz_l$$

$$\approx \sum_{i=1}^{N} \alpha_i f(\mathbf{z_i}), \quad (3.10)$$

where the region of integration, D, is a k-sphere of radius h. As a first step make a k-spherical transformation

$$\begin{aligned}
z_1 &= r\sin\varphi_1 \sin\varphi_2 \ldots \sin\varphi_{k-1},\\
z_2 &= r\cos\varphi_1 \sin\varphi_2 \ldots \sin\varphi_{k-1},\\
z_3 &= r\cos\varphi_2 \sin\varphi_3 \ldots \sin\varphi_{k-1},\\
&\ldots\\
z_{k-1} &= r\cos\varphi_{k-2} \sin\varphi_{k-1},\\
z_k &= r\cos\varphi_{k-1}.
\end{aligned}$$

The Jacobian of this transformation is

$$J(\varphi_1,\varphi_2,\ldots,\varphi_{k-1},r) = r^{k-1}\sin\varphi_2 \sin^2\varphi_3 \ldots \sin^{k-2}\varphi_{k-1}.$$

The set D is now defined by $0 \leq \varphi_1 \leq 2\pi$, $0 \leq \varphi_i \leq \pi$, where $i = 2, 3, \ldots, k-1$, and $0 \leq r \leq h$.

The rule to be constructed will have the easily computable form

$$I = \int_0^h \int_0^\pi \ldots \int_0^\pi \int_0^{2\pi} J f_1(\varphi_1, \varphi_2, \ldots, \varphi_{k-1}, r)\, d\varphi_1\, d\varphi_2 \ldots\, d\varphi_{k-1}\, dr$$
$$\approx \sum_{g,h,\ldots,i,j} \alpha_{g,h,\ldots,i,j} f_1(\varphi_{1,g}, \varphi_{2,h}, \ldots, \varphi_{k-1,i}, r_j), \quad (3.11)$$

where we will denote $f_1(\cdot)$ as

$$f_1(\varphi_1, \varphi_2, \ldots, \varphi_{k-1}, r) = \left(\frac{1}{2\pi}\right)^{(k-1)/2} \phi(r)$$
$$\times \prod_{m \in \{p-k+1,\ldots,p\}} \Phi(\min\{\{r\cos\varphi_{l-1}\sin\varphi_l \ldots \sin\varphi_{k-1}\}_{l=3,\ldots,k},$$
$$r\sin\varphi_1\sin\varphi_2\ldots\sin\varphi_{k-1}, r\cos\varphi_1\sin\varphi_2\ldots\sin\varphi_{k-1}\})$$

denoted for convenience. We want an approximation of precision K, i.e., it is exact for all functions $f_1(x_1, \ldots, x_p)$ of at most Kth degree and is not exact for some $(K+1)$st degree polynomial.

The quadrature rule suggested by Mustard (1964) for the integral over a k-dimensional spherical shell, leads to the approximation for the coverage probability that has a remarkably friendly computational form. The next result gives the form of the lower bound and also addresses the bounds accuracy.

Proposition 2. *The approximation for the lower bound of coverage probability*

$$P\left(\sum_{i=1}^{k}(X_{(i)} - \theta_{(i)})^2 \leq h^2\right) \geq \binom{p}{k} \sum_{g,h,\ldots,i,j} \alpha_{g,h,\ldots,i,j} f_1(\varphi_{1,g}, \varphi_{2,h}, \ldots, \varphi_{k-1,i}, r_j) \quad (3.12)$$

has precision $K = 4m + 3$, $m = 0, 1, 2\ldots$. Here

(a) the evaluation points are taken at all intersection points of the surfaces

$$\begin{aligned}
\varphi_1 &= \varphi_{1g} \qquad g = 1, 2, \ldots, 4(m+1), \\
\varphi_2 &= \varphi_{2h} \qquad h = 1, 2, \ldots, 2m+2, \\
&\ldots \\
\varphi_{k-1} &= \varphi_{k-1,i} \qquad i = 1, 2, \ldots, 2m+2, \\
r &= r_j \qquad j = 1, \ldots, m+1.
\end{aligned}$$

(b) the weight coefficients are given by $\alpha_{gh\ldots ij} = a_g b_{2,h} \ldots b_{k-1,i} c_j$*, where*

$$a_g = \frac{\pi}{2(m+1)}, \qquad g = 1, 2, \ldots, 4(m+1), \tag{3.13}$$

$$b_{\nu,i} = \frac{\left(\left(2m+2+\frac{\nu-1}{2}\right)!\right)^2}{(2m+2)!(2m+\nu+1)!} \cdot \frac{2^\nu}{(1-y_i^2)\left(Q'^{(\nu-1)/2}_{2m+2}(y_i)\right)^2}, \tag{3.14}$$

$$\nu = l - 1 = 1, \ldots, k-2, \; i = 1, 2, \ldots, 2m+2, \tag{3.15}$$

$$c_j = \frac{2^{k/2-1}}{\left(m+1+\frac{k}{2}\right)^2 \frac{r_j^2}{h^2}\left(1-\frac{r_j^2}{h^2}\right)\left[(m+1)\,{}_2F_1\left(-m, m+\frac{k}{2}+2; 2; 1-\frac{r_j^2}{h^2}\right)\right]^2}. \tag{3.16}$$

Here $y_i = cos\varphi_{\nu i}$ *are the* $2m+2$ *zeroes of the orthogonal polynomials* $Q^{(\nu-1)/2}_{2m+2}(y)$*, where*

$$Q_p^\alpha = \binom{p+\alpha}{p} G_p\left(2\alpha+1, \alpha+1, \frac{1}{2}(1-y)\right), \tag{3.17}$$

where $G_p(a, b, z)$ *are the Jacobi polynomials and* $Q_p^{(\alpha,\alpha)} \equiv Q_p^{(\alpha)}$*. The* r_j^2 *are the* $m+1$ *zeroes of the polynomial*

$${}_2F_1\left(-(m+1), m+\frac{k}{2}+1; 1; 1-\frac{r^2}{h^2}\right) \tag{3.18}$$

of degree $m+1$ *in* r^2*.*

Proof. We first rewrite integral I in the following form

$$I = \left(\frac{1}{2\pi}\right)^{(k-1)/2} \int_0^h r^{k-1}\phi(r) \int_0^\pi \sin^{k-2}\varphi_{k-1} \int_0^\pi \dots \int_0^\pi \sin\varphi_2$$
$$\times \int_0^{2\pi} \prod_{m\in\{p-k+1,\dots,p\}} \Phi\left(\min_{l=\{3,\dots,k\}} [\{r\cos\varphi_{l-1}\sin\varphi_l \dots \sin\varphi_{k-1} + \theta_l\};\right.$$
$$r\cos\varphi_1\sin\varphi_2\dots\sin\varphi_{k-1} + \theta_2; r\sin\varphi_1\sin\varphi_2\dots\sin\varphi_{k-1}$$
$$\left. +\theta_1] - \theta_m\right)\, d\varphi_1\, d\varphi_2 \dots\, d\varphi_{k-1}\, dr. \tag{3.19}$$

Now define functions $f_2, \dots, f_k$ as follows:

$$f_2(\varphi_2, \dots, \varphi_{k-1}, r) = \int_0^{2\pi} f_1(\varphi_1, \dots, \varphi_{k-1}, r)\, d\varphi_1, \tag{3.20}$$
$$f_l(\varphi_l, \dots, \varphi_{k-1}, 2) = \int_0^\pi \sin^{l-2}\varphi_{l-1} f_{l-1}(\varphi_{l-1}, \dots, \varphi_{k-1}, r)\, d\varphi_{l-1}, \tag{3.21}$$
$$(l = 2, 3, \dots, k-1),$$
$$f_k(r) = \int_0^\pi r^{k-1}\sin^{k-2}\varphi_{k-1} f_{k-1}(\varphi_{k-1}, r)\, d\varphi_{k-1}, \tag{3.22}$$

then it follows that

$$I = \int_0^h r^{k-1} f_k(r)\, dr.$$

We want to find the collection of quadrature rules of the form

$$\int_0^{2\pi} f_1\, d\varphi_1 \approx \sum_{g=1}^{p_1} a_g f_1(\varphi_{1,g}, \varphi_2, \dots, \varphi_{k-1}, 2),$$
$$\int_0^\pi \sin^{l-1}\varphi_l f_l\, d\varphi_l \approx \sum_{i=1}^{p_l} b_{li} f_l(\varphi_{l,i}, \varphi_{l+1}, \dots, \varphi_{k-1}, r) \tag{3.23}$$
$$(l = 2, 3, \dots, k-1),$$
$$\int_0^h r^{k-1} f_k(r)\, dr \approx \sum_{j=1}^{p_k} c_j f_k(r_j).$$

Consider the integral of f_2 in (3.20),

$$f_2(\varphi_2, \ldots, \varphi_{k-1}, r) =$$
$$= \int_0^{2\pi} \prod_{m \in \{p-k+1,\ldots,p\}} \Phi\left(\min_{l=\{2,\ldots,k\}} [\{r \cos\varphi_{l-1} \sin\varphi_l \ldots \sin\varphi_{k-1} + \theta_l\}; \right.$$
$$\left. r \sin\varphi_1 \sin\varphi_2 \ldots \sin\varphi_{k-1} + \theta_1] - \theta_m \right) d\varphi_1.$$

In order to evaluate this integral we need to consider three separate cases according to what minimum inside the normal cdf could be:

Case 1. In this case $\min = r \sin\varphi_1 \sin\varphi_2 \ldots \sin\varphi_{k-1} + \theta_1$ $(= z_1 + \theta_1$ in the original notation). Then

$$I_1 = \int_0^{2\pi} \prod_{m \in \{p-k+1,\ldots,p\}} \Phi\left(r \sin\varphi_1 \sin\varphi_2 \ldots \sin\varphi_{k-1} + \theta_1 - \theta_m\right) d\varphi_1$$
$$\approx \sum_{g=1}^{p_1} a_g \prod_{m \in \{p-k+1,\ldots,p\}} \Phi\left(r \sin\varphi_{1,g} \sin\varphi_2 \ldots \sin\varphi_{k-1} + \theta_1 - \theta_m\right),$$

where

$$a_g = \frac{2\pi}{p_1} = \frac{2\pi}{4(m+1)}$$
$$\varphi_{1,g} = \frac{2\pi g}{p_1} = \frac{2\pi g}{4(m+1)}, \qquad g = 1, 2, \ldots, 4(m+1).$$

Then the number of evaluation points, p_1, is chosen to guarantee that approximation has precision K. In this case

$$I = \left(\frac{1}{2\pi}\right)^{(k-1)/2} \sum_{g=1}^{4(m+1)} \frac{2\pi}{4(m+1)} \int_0^h r^{k-1} \phi(r) \int_0^\pi \sin^{k-2}\varphi_{k-1} \int_0^\pi \ldots \int_0^\pi \sin\varphi_2$$
$$\times \prod_{m \in \{p-k+1,\ldots,p\}} \Phi\left(r \sin\left(\frac{2\pi g}{4(m+1)}\right) \sin\varphi_2 \ldots \sin\varphi_{k-1} + \theta_1 - \theta_m\right)$$
$$d\varphi_2 \ldots d\varphi_{k-1}\, dr. \tag{3.24}$$

Following Mustard (1964) we can write

$$\int_0^{\pi} \sin^{\nu}\varphi_l \prod_{m\in\{p-k+1,\ldots,p\}} \Phi\left(r\sin\left(\frac{2\pi g}{4(m+1)}\right)\sin\varphi_l \sin\varphi_{l+1}\ldots\sin\varphi_{k-1} \right.$$
$$\left. +\theta_1-\theta_m\right)\, d\varphi_t$$
$$\approx \sum_{i=1}^{p_t} b_{\nu,i} \prod_{m\in\{p-k+1,\ldots,p\}} \Phi\left(r\sin\left(\frac{2\pi g}{4(m+1)}\right)\sin\varphi_{l,i} \sin\varphi_{l+1}\ldots\sin\varphi_{k-1} \right.$$
$$\left. +\theta_1-\theta_m\right),$$

for $l=2,\ldots,k-1$, $\nu=l-1$ denoted for convenience, where $b_{\nu,i}$ are given in (3.15) and the $\cos\varphi_{p-k+\nu,i}$ $(i=1,2,\ldots,p_{\nu+1})$ terms are the $2m+2$ zeros of the orthogonal polynomials $Q_{2m+2}^{(\nu-1)/2}(y)$, where $Q_p^{\alpha}(y)$ is given in (3.17).

Thus, using the relation between the Jacobi polynomials and the hypergeometric function, we can write

$$Q_{2m+2}^{(\nu-1)/2}(y_i) = \binom{2m+2+\frac{\nu-1}{2}}{2m+2} G_{2m+2}\left(\nu, \frac{\nu+1}{2}; \frac{1}{2}(1-y_i)\right)$$
$$= \binom{2m+2+\frac{\nu-1}{2}}{2m+2} {}_2F_1\left(-(2m+2), 2m+\nu+2; \frac{\nu+1}{2}; \frac{1}{2}(1-y_i)\right) \tag{3.25}$$

and

$$Q'^{(\nu-1)/2}_{2m+2}(y_i) = \binom{2m+2+\frac{\nu-1}{2}}{2m+2} \frac{(2m+2)(2m+\nu+2)}{\nu+1} \tag{3.26}$$
$$\times {}_2F_1\left(-(2m+1), 2m+\nu+3; \frac{\nu+3}{2}; \frac{1}{2}(1-y_i)\right), \tag{3.27}$$

since

$$\frac{d_2F_1(a,b;c;z)}{dz} = \frac{ab}{c} {}_2F_1\left(a+1, b+1; c+1; z\right).$$

The number of evaluation points $p_{\nu+1}=2m+2$ is chosen again to guarantee that the approximation has precision $K=4m+3$.

Substituting in I_1 we get

$$
\begin{aligned}
I = \left(\frac{1}{2\pi}\right)^{(k-1)/2} & \sum_{g=1}^{4(m+1)} \frac{2\pi}{4(m+1)} \int_0^h r^{k-1}\phi(r) \sum_{h,\ldots,i}^{2m+2} b_{2,h}\ldots b_{k-1,i} \\
& \times \prod_{m\in\{p-k+1,\ldots,p\}} \Phi\left(r\sin\left(\frac{2\pi g}{4(m+1)}\right)\sin\varphi_{2,h}\sin\varphi_{3,i}\ldots\sin\varphi_{k-1,j}\right. \\
& \left.+\theta_1-\theta_m\right)\,dr \\
= \left(\frac{1}{2\pi}\right)^{(k-1)/2} & \sum_{g=1}^{4(m+1)} \sum_{j=1}^{m+1} \sum_{h,\ldots,i}^{2m+2} \frac{\pi}{2(m+1)} b_{2,h}\ldots b_{k-1,i}c_j\phi(r_j) \\
& \times \prod_{m\in\{p-k+1,\ldots,p\}} \Phi\left(r_j\sin\left(\frac{\pi g}{2(m+1)}\right)\sin\varphi_{2,h}\sin\varphi_{3,i}\ldots\sin\varphi_{k-1,j}\right. \\
& \left.+\theta_1-\theta_m\right),
\end{aligned}
$$

where the values of c_j, r_j, are similar to the values in Mustard (1964). The difference is that integral in Mustard (1964) is over the interval $[0,1]$, here the integral is over the interval $[0,h]$. Thus, we need to do appropriate transformation to get the proper estimates of c_j and r_j.

The integral we need to estimate has the following form

$$I = \int_0^h r^{k-1}g(r^2)\,dr. \tag{3.28}$$

Using quadrature rules, we want to get the approximation

$$\int_0^h r^{k-1}g(r^2)\,dr \approx \sum_{j=1}^{g} c_j g(r_j^2),$$

or putting $t = r^2$

$$\int_0^{h^2} t^{k/2-1}g(t)\,dt \approx \sum_{j=1}^{g} 2c_j g(t_j).$$

The following calculations to get the approximation of the interval above are based on Section 7.3 in Krylov & Stroud (2006), where they consider getting a quadrature approximation of the integral $\int_a^b (b-x)^\alpha (x-a)^\beta g(x)\,dx$ which corresponds to our case with $a=0$, $b=h^2$, $\alpha=0$, and $\beta=\frac{k}{2}-1$. To construct

quadrature formulas for the approximation of the integral it is common to transform it to the integral over the interval $[-1,1]$ by the linear transformation

$$\begin{aligned} t &= \frac{1}{2}\left[0 + h^2 + s(h^2 - 0)\right] = \frac{1}{2}h^2(1+s), \qquad -1 \le s \le 1, \\ s &= \frac{2t}{h^2} - 1 = \frac{2r^2}{h^2} - 1, \\ dt &= \frac{h^2}{2}\, ds. \end{aligned} \tag{3.29}$$

Substituting this transformation into (3.28), we get

$$\begin{aligned} I = \int_0^{h^2} t^{k/2-1} g(t)\, dt &= \int_{-1}^{1} \left(\frac{h^2}{2}(1+s)\right)^{k/2-1} g\left(\frac{h^2}{2}(1+s)\right) \frac{h^2}{2}\, ds \\ &= \left(\frac{h}{\sqrt{2}}\right)^k \int_{-1}^{1} (1+s)^{k/2-1} g\left(\frac{h^2}{2}(1+s)\right) ds. \end{aligned} \tag{3.30}$$

The orthogonal system of polynomials which correspond to the segment $[-1,1]$ and the weight function $(1-s)^\alpha(1+s)^\beta = (1+s)^{k/2-1}$ is the system of Jacobi polynomials

$$P_n^{(\alpha,\beta)}(s) = P_n^{\left(0,\frac{k}{2}-1\right)}(s), \qquad n = 0,1,2,\ldots\,.$$

The quadrature formula with n nodes is then

$$\int_{-1}^{1} (1+s)^{k/2-1} g\left(\frac{h^2}{2}(1+s)\right) ds \approx \sum_{j=1}^{n} 2c_j g\left(\frac{h^2}{2}(1+s_j)\right), \tag{3.31}$$

which has the highest degree of precision $2n-1$, when nodes s_j are the roots of the Jacobi polynomial of degree n. That is we need to solve

$$P_n^{(\alpha,\beta)}(s_j) = P_n^{\left(0,\frac{k}{2}-1\right)}(s_j) \equiv 0. \tag{3.32}$$

Using the connection between Jacobi polynomials and hypergeometric functions

$$P_n^{(\alpha,\beta)}(z) = \binom{n+\alpha}{n} {}_2F_1\left(-n, n+\alpha+\beta+1; \alpha+1; \frac{1}{2}(1-z)\right),$$

we can rewrite the equation (3.32) as

$$
\begin{aligned}
{}_2F_1\left(-n, n+\frac{k}{2}; 1; \frac{1}{2}(1-s_j)\right) &= 0, \\
{}_2F_1\left(-(m+1), m+\frac{k}{2}+1; 1; \frac{1}{2}(1-s_j)\right) &= 0.
\end{aligned}
$$

Replacing s_j with (3.29), we get

$$
{}_2F_1\left(-(m+1), m+\frac{k}{2}+1; 1; 1-\frac{r^2}{h^2}\right) = 0. \tag{3.33}
$$

The weights c_j are given in (7.3.4) in Krylov & Stroud (2006):

$$
\begin{aligned}
2c_j &= \frac{2^{\alpha+\beta+1}\Gamma(\alpha+n+1)\Gamma(\beta+n+1)}{n!\Gamma(\alpha+\beta+n+1)(1-s_j^2)\left[P'^{(\alpha,\beta)}_n(s_j)\right]^2} \\
&= \frac{2^{k/2}\Gamma(n+1)\Gamma\left(n+\frac{k}{2}\right)}{n!\Gamma\left(n+\frac{k}{2}\right)(1-s_j^2)\left[P'^{\left(0,\frac{k}{2}-1\right)}_n(s_j)\right]^2} \\
&= \frac{2^{k/2}}{(1-s_j^2)\left[P'^{\left(0,\frac{k}{2}-1\right)}_n(s_j)\right]^2}.
\end{aligned} \tag{3.34}
$$

The derivative of Jacobi polynomial is given by

$$
\begin{aligned}
\frac{d}{dx}P^{(\alpha,\beta)}_n(x) &= \frac{1}{2}(n+\alpha+\beta+1)P^{(\alpha+1,\beta+1)}_{n-1}(x) \\
\frac{d}{dx}P^{\left(0,\frac{k}{2}-1\right)}_n(x) &= \frac{1}{2}\left(n+\frac{k}{2}\right)P^{\left(1,\frac{k}{2}\right)}_{n-1}(x).
\end{aligned}
$$

Taking this into to account, we get

$$
\begin{aligned}
2c_j &= \frac{2^{k/2+2}}{\left(n+\frac{k}{2}\right)^2(1-s_j^2)\left[P^{\left(1,\frac{k}{2}\right)}_{n-1}(s_j)\right]^2} \\
&= \frac{2^{k/2+2}}{\left(n+\frac{k}{2}\right)^2(1-s_j^2)\left[n\,{}_2F_1\left(-(n-1), n+\frac{k}{2}+1; 2; \frac{1}{2}(1-s_j)\right)\right]^2} \\
&= \frac{2^{k/2+2}}{\left(m+1+\frac{k}{2}\right)^2(1-s_j^2)\left[(m+1)\,{}_2F_1\left(-m, m+\frac{k}{2}+2; 2; \frac{1}{2}(1-s_j)\right)\right]^2}.
\end{aligned}
$$

Again, substitute r_j using (3.29)

$$
c_j = \frac{2^{k/2-1}}{\left(m+1+\frac{k}{2}\right)^2\frac{r^2}{h^2}\left(1-\frac{r^2}{h^2}\right)\left[(m+1)\,{}_2F_1\left(-m, m+\frac{k}{2}+2; 2; 1-\frac{r^2}{h^2}\right)\right]^2}. \tag{3.35}
$$

We get the following quadrature formula (here n is taken to be $n = m + 1$ which provides the precision of the approximation equal to $4m + 3$)

$$I = \int_0^h r^{k-1} f_k(r)\, dr \approx \left(\frac{h}{\sqrt{2}}\right)^k \sum_{j=1}^{m+1} c_j f_k\left(r_j\right). \tag{3.36}$$

Case 2. In this case $\min = r\cos\varphi_1 \sin\varphi_2 \ldots \sin\varphi_{k-1} + \theta_2$ $(= z_2 + \theta_2$ in the original notation). Then

$$\begin{aligned} I_1 &= \int_0^{2\pi} \prod_{m\in\{p-k+1,\ldots,p\}} \Phi\left(r\cos\varphi_1 \sin\varphi_2 \ldots \sin\varphi_{k-1} + \theta_2 - \theta_m\right)\, d\varphi_1 \\ &\approx \sum_{g=1}^{4(m+1)} \frac{2\pi}{4(m+1)} \prod_{m\in\{p-k+1,\ldots,p\}} \Phi\left(r\cos\left(\frac{2\pi g}{4(m+1)}\right)\sin\varphi_2 \ldots \sin\varphi_{k-1}\right. \\ &\qquad \left. + \theta_2 - \theta_m\right). \end{aligned} \tag{3.37}$$

As in the Case 1,

$$\begin{aligned} I = &\left(\frac{1}{2\pi}\right)^{(k-1)/2} \left(\frac{h}{\sqrt{2}}\right)^k \sum_{g=1}^{4(m+1)} \sum_{h,\ldots,i}^{2m+2} \frac{\pi}{2(m+1)} b_{2,h} \ldots b_{k-1,i} c_j \phi(r_j) \\ &\times \prod_{m\in\{p-k+1,\ldots,p\}} \Phi\left(r_j \cos\left(\frac{\pi g}{2(m+1)}\right)\sin\varphi_{2,h}\sin\varphi_{3,i}\ldots\sin\varphi_{k-1,j} + \theta_2 - \theta_m\right). \end{aligned} \tag{3.38}$$

Case 3. In this case $\min = \min\limits_{l\in\{3,\ldots,k\}} \left\{r\cos\varphi_{l-1}\sin\varphi_l \ldots \sin\varphi_{k-1} + \theta_l\right\}$ and

$$I_1 = 2\pi \prod_{m\in\{p-k+1,\ldots,p\}} \Phi\left(\min_{l\in\{3,\ldots,k\}} \left\{r\cos\varphi_{l-1}\sin\varphi_l \ldots \sin\varphi_{k-1} + \theta_l\right\} - \theta_m\right).$$

Then the integral of interest is

$$\begin{aligned} I = &\left(\frac{1}{2\pi}\right)^{(k-1)/2} 2\pi \int_0^h r^{k-1}\phi(r) \int_0^\pi \sin^{k-2}\varphi_{k-1} \int_0^\pi \ldots \int_0^\pi \sin\varphi_2 \\ &\times \prod_{m\in\{p-k+1,\ldots,p\}} \Phi\left(\min_{l\in\{3,\ldots,p\}} \left\{r\cos\varphi_{l-1}\sin\varphi_l \ldots \sin\varphi_{k-1} + \theta_l\right\} - \theta_m\right)\, d\varphi_2 \\ &d\varphi_3 \ldots\, d\varphi_{k-1}\, dr. \end{aligned} \tag{3.39}$$

Again, by considering separate cases for what min could be we have

Case 3 (a). If $\min = r\cos\varphi_2 \sin\varphi_3 \ldots \sin\varphi_{k-1} + \theta_3 \equiv z_3 - \theta_3$, then

$$\begin{aligned} I_2 &= \int_0^\pi \sin\varphi_2 \prod_{m\in\{p-k+1,\ldots,p\}} \Phi\left(\min_{l\in\{3,\ldots,k\}} \{r\cos\varphi_{l-1}\sin\varphi_l \ldots \sin\varphi_{k-1} \right. \\ &\qquad \left. +\theta_l\} - \theta_m\right)\, d\varphi_2 \\ &= \int_0^\pi \sin\varphi_2 \prod_{m\in\{p-k+1,\ldots,p\}} \Phi\left(r\cos\varphi_2\sin\varphi_3\ldots\sin\varphi_{k-1} + \theta_3 - \theta_m\right)\, d\varphi_2 \\ &\approx \sum_i^{2m+2} b_{2,i} \prod_{m\in\{p-k+1,\ldots,p\}} \Phi\left(r\cos\varphi_{2,i}\sin\varphi_3\ldots\sin\varphi_{k-1} + \theta_3 - \theta_m\right). \end{aligned}$$

Therefore,

$$\begin{aligned} I &= \left(\frac{1}{2\pi}\right)^{(k-1)/2} \left(\frac{h}{\sqrt{2}}\right)^k 2\pi \sum_{h,\ldots,i}^{2m+2} b_{2,h}\ldots b_{k-1,i} c_j \phi(r_j) \\ &\quad \times \prod_{m\in\{p-k+1,\ldots,p\}} \Phi\left(r_j\cos\varphi_{2,h}\sin\varphi_{3,l}\ldots\sin\varphi_{k-1,i} + \theta_3 - \theta_m\right). \end{aligned} \tag{3.40}$$

Case 3 (b). If $\min = \min\limits_{l\in\{4,\ldots,k-1\}} \{r\cos\varphi_{l-1}\sin\varphi_l\ldots\sin\varphi_{k-1} + \theta_l\} \equiv z_l - \theta_l$,

$$\begin{aligned} I_2 &= \int_0^\pi \sin\varphi_2 \prod_{m\in\{p-k+1,\ldots,p\}} \Phi\left(\min_{l\in\{4,\ldots,k\}} \{r\cos\varphi_{l-1}\sin\varphi_l\ldots\sin\varphi_{k-1} \right. \\ &\qquad \left. +\theta_l\} - \theta_m\right)\, d\varphi_2 \\ &= 2 \prod_{m\in\{p-k+1,\ldots,p\}} \Phi\left(\min_{l\in\{4,\ldots,k\}} \{r\cos\varphi_{l-1}\sin\varphi_l\ldots\sin\varphi_{k-1} + \theta_l\} - \theta_m\right), \end{aligned}$$

and

$$\begin{aligned} I &= \left(\frac{1}{2\pi}\right)^{k/2} \left(\frac{h}{\sqrt{2}}\right)^k 2\pi\cdot 2 \int_0^h r^{k-1}\phi(r) \int_0^\pi \sin^{k-2}\phi_{k-1}\ldots\int_0^\pi \sin^2\varphi_3 \\ &\quad \times \prod_{m\in\{p-k+1,\ldots,p\}} \Phi\left(\min_{l\in\{4,\ldots,k\}} \{r\cos\varphi_{l-1}\sin\varphi_l\ldots\sin\varphi_{k-1} + \theta_l\} - \theta_m\right) \\ &\quad d\varphi_3\, d\varphi_4 \ldots\, d\varphi_{k-1}\, dr. \end{aligned} \tag{3.41}$$

If we keep moving in the similar fashion, we will get the following approxima-

tion for I_1, then

$$I = (2\pi)^{(1-k)/2} \left(\frac{h}{\sqrt{2}}\right)^k \left(\int_0^\pi \sin\varphi_2 \, d\varphi_2\right) \ldots \left(\int_0^\pi \sin^{k-2}\varphi_{k-1} \, d\varphi_{k-1}\right)$$
$$\times \sum_{h,\ldots,i}^{2m+2} b_{t-1,h} \ldots b_{k-1,i} c_j \phi(r_j)$$
$$\times \prod_{m\in\{p-k+1,\ldots,p\}} \Phi\left(r_j \cos\varphi_{t-1,h} \sin\varphi_{t+2,i} \ldots \sin\varphi_{k-1,j} + \theta_t - \theta_m\right), \qquad (3.42)$$

where $t \in \{3, \ldots, k\}$ corresponds to which $z_t - \theta_t$ was the minimum.

This probability is minimized when $\theta_t = \theta_m = 0$ for all $t = 3, \ldots, k$ and $m = p - k + 1, \ldots, p$. □

Combining the various conclusions above, we can get the following

Theorem 4. *The coverage probability of the confidence set for the selected normal means, corresponding to the means of the populations with the largest sample means, can be approximated with a precision of* $4m + 3$ *by*

(i) If $\min_{l\in\{1,\ldots,k\}} \{z_l + \theta_l\} = z_1 + \theta_1$, *then*

$$P\left(\|\boldsymbol{\theta}_s - \mathbf{X}_s\|^2 \le h^2\right) \approx \binom{p}{k} \left(\frac{1}{2\pi}\right)^{(k-1)/2} \left(\frac{h}{\sqrt{2}}\right)^k$$
$$\times \sum_{g=1}^{4(m+1)} \sum_{h,i,\ldots,q,j}^{2m+2} \sum_{j=1}^{m+1} \frac{\pi}{2(m+1)} b_{2,h} \ldots b_{k-1,q} c_j \phi\left(r_j\right))$$
$$\times \prod_{m\in I_j^c} \Phi\left(r_j \sin\left(\frac{\pi g}{2(m+1)}\right) \sin\varphi_{2,h} \sin\varphi_{3,i} \ldots \sin\varphi_{k-1,q} + \theta_2 - \theta_m\right),$$

where the evaluation points r_j *are given by* (3.33) *and weights* c_j *by* (3.35).

(ii) If $\min_{l\in\{1,\ldots,k\}}\{z_l+\theta_l\}=z_2+\theta_2$, *then*

$$P\left(\|\boldsymbol{\theta}_s-\mathbf{X}_s\|^2\le h^2\right)\approx\binom{p}{k}\left(\frac{1}{2\pi}\right)^{(k-1)/2}\left(\frac{h}{\sqrt{2}}\right)^k$$
$$\times\sum_{g=1}^{4(m+1)}\sum_{h,i,\ldots,q,j}^{2m+2}\sum_{j=1}^{m+1}\frac{\pi}{2(m+1)}b_{2,h}\ldots b_{k-1,q}c_j\phi\left(r_j\right)$$
$$\times\prod_{m\in I_j^c}\Phi\left(r_j\cos\left(\frac{\pi g}{2(m+1)}\right)\sin\varphi_{2,h}\sin\varphi_{3,i}\ldots\sin\varphi_{k-1,q}+\theta_2-\theta_m\right).$$

(iii) If $\min_{l\in\{1,\ldots,k\}}\{z_l+\theta_l\}=z_t+\theta_t$, *where* $t\in\{3,\ldots,k\}$, *then*

$$P\left(\|\boldsymbol{\theta}_s-\mathbf{X}_s\|^2\le h^2\right)\approx\binom{p}{k}\left(\frac{1}{2\pi}\right)^{(k-1)/2}\left(\frac{h}{\sqrt{2}}\right)^k\left(\int_0^\pi\sin\varphi_2\,d\varphi_2\right)$$
$$\times\cdots\times\left(\int_0^\pi\sin^{k-2}\varphi_{k-1}\,d\varphi_{k-1}\right)\sum_{h,i,\ldots,q}^{2m+2}\sum_{j=1}^{m+1}b_{t-1,h}\ldots b_{k-1,q}c_j\phi\left(r_j\right)$$
$$\times\prod_{m\in I_j^c}\Phi\left(r_j\cos\varphi_{t-1,h}\sin\varphi_{t+2,i}\ldots\sin\varphi_{k-1,q}+\theta_t-\theta_m\right).$$

In practice we will not know the minimum of $\{z_l-\theta_l\}_{l=1}^k$ since θ_l is unknown. The result of Theorem should be used to simplify simulations. For real data applications Proposition 3 should be used directly.

Note that the above approximations are minimized when $\theta_t=\theta_m$ for all $m\in I_j^c$. Furthermore, the integrals involving sin in the Theorem 3.2.1 above are constants and are

$$\int_0^\pi\sin\varphi_2\,d\varphi_2=2$$
$$\int_0^\pi\sin^2\varphi_3\,d\varphi_3=\frac{\pi}{2}$$
$$\ldots$$
$$\int_0^\pi\sin^{k-2}\varphi_{k-1}\,d\varphi_{k-1}=-\frac{\sin^{k-3}\varphi_{k-1}\cos\varphi_{k-1}}{k-2}\Big|_0^\pi+\frac{k-3}{k-2}\int_0^\pi\sin^{k-4}\varphi_{k-1}\,d\varphi$$
$$=\frac{k-3}{k-2}\int_0^\pi\sin^{k-4}\varphi_{k-1}\,d\varphi\qquad\text{for } k\ge 4.$$

To find a confidence set for the selected mean with guaranteed coverage probability of $1-\alpha$, we need to solve the following for h

$$\min_{\boldsymbol{\theta}} \left\{ \text{Approximation of } P\left(\|\boldsymbol{\theta}_s - \mathbf{X}_s\|^2 \leq h^2\right) \text{ from Proposition 3} \right\} \geq 1-\alpha. \tag{3.43}$$

Numerical check of the above approximation

In this section we will evaluate the approximation for the coverage probability of the confidence set for the selected means derived above.

In general, the weights $b_{l,h}, \ldots, b_{l,i}$, c_j and the evaluation points t_j depend on m, which defines the desired precision, p, total number of populations, and k, the number of selected populations. Mustard (1964) provided tabulated values for some configurations, we copy some of them below.

Case general p and $k = 2$

For general number of populations p and number of selected populations $k = 2$, the coverage probability is

$$\begin{aligned} & P\left(\sum_{i=1}^{2} \left(X_{(i)} - \theta_{(i)}\right)^2 \leq h^2\right) \\ &= \binom{p}{k} \left(\frac{h}{\sqrt{2}}\right)^2 \iint\limits_{\sum_{q\in I_j} z_q^2 \leq h^2} \prod_{m \in I_j^c} \Phi\left(\min_{l\in I_j}\{z_l + \theta_l\} - \theta_m\right) \prod_{l\in I_j} \phi(z_l)\, dz_l \\ &= \binom{p}{k} \frac{h^2}{2} \iint\limits_{z_1^2+z_2^2 \leq h^2} \prod_{m\in\{3,p\}} \Phi\left(\min\{z_1+\theta_1, z_2+\theta_2\} - \theta_m\right) \phi(z_1)\phi(z_2)\, dz_1\, dz_2. \end{aligned} \tag{3.44}$$

Making the transformation $z_1 = r\cos\varphi$ and $z_2 = r\sin\varphi$ we can get a lower bound for this expression by taking $\theta_1 = \theta_2 = \theta_m$, then (3.44) is greater than

$$\binom{p}{k} \frac{h^2}{2} \left(\frac{1}{2\pi}\right)^{1/2} \int_0^h \int_0^{2\pi} r\phi(r)\Phi^{p-2}\left(r\min\{\cos\varphi, \sin\varphi\}\right)\, dr\, d\varphi. \tag{3.45}$$

By splitting the integral over ϕ into integrals over intervals where we know for each value of $\min\{\cos\varphi, \sin\varphi\}$, (3.45) equals

$$\binom{p}{k} \frac{h^2}{2\sqrt{2\pi}} \int_0^r \left[\int_0^{\pi/4} \Phi^{p-2}(r\sin\varphi)\, d\varphi + \int_{\pi/4}^{5\pi/4} \Phi^{p-2}(r\cos\varphi)\, d\varphi + \int_{5\pi/4}^{2\pi} \Phi^{p-2}(r\sin\varphi)\, d\varphi \right] dr. \tag{3.46}$$

For instance, by equating (3.46) to $1-\alpha$ it could be easily solved numerically to provide $h = 2.34$ for $p = 3$.

Part of the motivation for using quadrature rules, as in Section 2.3.2, we obtain an approximation of the coverage probability is to split up the interval $[0, 2\pi]$ according to the $\min_{l\in\{1:k\}}\{z_l\} = r\min\{\{\cos\varphi_{l-1}\sin\varphi_l \ldots \sin\varphi_{k-1}\}_{l\in\{2,\ldots,k\}}, \sin\varphi_1 \sin\varphi_2 \ldots \sin\varphi_{k-1}\}$ terms.

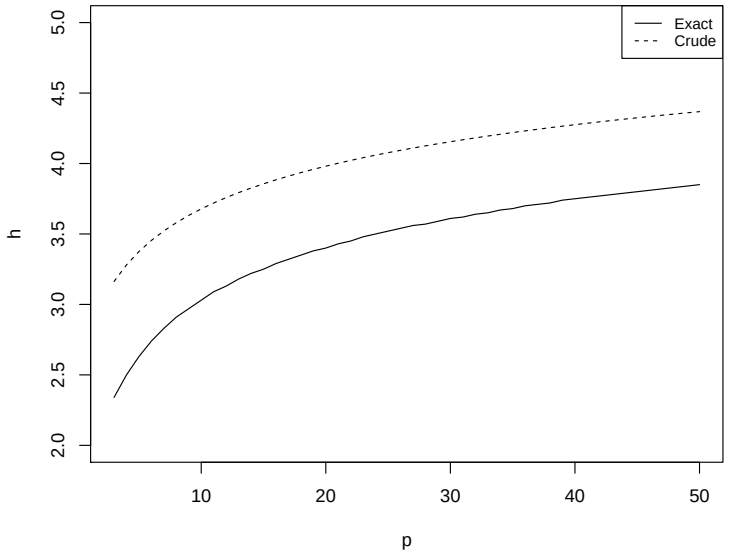


Figure 3.3: Radius of the confidence set for the lower bound (3.46) (exact) and approximation (3.9) (crude) for $k = 2$.

First, consider the case $m = 0$, then the coverage probability is

$$
\begin{aligned}
&P\left(\sum_{i=1}^{k}(X_{(i)}-\theta_{(i)})^2 \le h^2\right) \\
&\quad\approx \binom{p}{k}\left(\frac{1}{2\pi}\right)^{(k-1)/2}\left(\frac{h}{\sqrt{2}}\right)^k \sum_{g=1}^{4(m+1)}\sum_{j=1}^{m+1}\frac{\pi}{2(m+1)}c_j\phi\left(r_j\right) \\
&\qquad\quad\times\prod_{m\in I_j^c}\Phi\left(r_j\sin\frac{\pi g}{2(m+1)}+\theta_{(2)}-\theta_{(m)}\right) \\
&\quad= \binom{p}{k}\left(\frac{1}{2\pi}\right)^{(k-1)/2}\left(\frac{h}{\sqrt{2}}\right)^k \sum_{g=1}^{4}\frac{\pi}{2}c_1\phi\left(r_1\right)\prod_{m\in I_j^c}\Phi\left(r_1\sin\frac{\pi g}{2}+\theta_{(2)}-\theta_{(m)}\right).
\end{aligned}
\tag{3.47}
$$

Estimation points r_j are the solutions of the equation (3.33)

$$
{}_2F_1\left(-(m+1), m+1+\frac{k}{2}; 1; 1-\frac{r^2}{h^2}\right) = 0. \tag{3.48}
$$

In this case

$$
{}_2F_1\left(-1, \frac{k}{2}+1; 1; 1-\frac{r^2}{h^2}\right) = 0. \tag{3.49}
$$

The solution to (3.49) is

$$
r_1^2 = \frac{k}{k+2}h^2. \tag{3.50}
$$

The weight c_1 is given by (3.35) as

$$
\begin{aligned}
c_1 &= \frac{2^{k/2-1}}{\left(m+1+\frac{k}{2}\right)^2\frac{r^2}{h^2}\left(1-\frac{r^2}{h^2}\right)\left[(m+1){}_2F_1\left(-m, m+\frac{k}{2}+2; 2; 1-\frac{r^2}{h^2}\right)\right]^2} \\
&= \frac{2^{k/2-1}}{\left(\frac{k}{2}+1\right)^2\frac{r^2}{h^2}\left(1-\frac{r^2}{h^2}\right)\left[{}_2F_1\left(0, \frac{k}{2}+2; 2; 1-\frac{r^2}{h^2}\right)\right]^2} \\
&= \frac{2^{k/2}}{k}.
\end{aligned}
\tag{3.51}
$$

Hence, putting all of these terms together, the approximation for coverage prob-

ability is given by

$$P\left(\sum_{i=1}^{k}(X_{(i)}-\theta_{(i)})^2 \le h^2\right) \approx \binom{p}{k}\left(\frac{1}{2\pi}\right)^{(k-1)/2}\left(\frac{h}{\sqrt{2}}\right)^k \frac{2^{k/2}}{k}\phi\left(h\sqrt{\frac{k}{k+2}}\right)$$
$$\times \sum_{g=1}^{4}\frac{\pi}{2}\prod_{m\in I_j^c}\Phi\left(h\sqrt{\frac{k}{k+2}}\sin\frac{\pi g}{2}+\theta_{(2)}-\theta_{(m)}\right)$$
$$\ge \binom{p}{k}\left(\frac{1}{2\pi}\right)^{(k-1)/2}\left(\frac{h}{\sqrt{2}}\right)^k \frac{2^{k/2}}{k}\phi\left(h\sqrt{\frac{k}{k+2}}\right)\sum_{g=1}^{4}\frac{\pi}{2}\Phi^{p-k}\left(h\sqrt{\frac{k}{k+2}}\sin\frac{\pi g}{2}\right).$$

Proposition 3. *The confidence set for selecting $k = 2$ means, corresponding to the populations with the largest sample means, is given by $C = \left\{\sum_{i=1}^{2}(X_{(i)}-\theta_{(i)})^2 \le h^{*2}\right\}$, where h^* is the solution of*

$$\binom{p}{2}\frac{h^2\sqrt{\pi}}{2\sqrt{2}}\phi\left(\frac{h}{\sqrt{2}}\right)\sum_{g=1}^{4}\Phi^{p-k}\left(\frac{h}{\sqrt{2}}\sin\frac{\pi g}{2}\right)=1-\alpha. \tag{3.52}$$

We have the following observations. First, for selecting $k = 2$ populations, we only need to consider Case 1 and Case 2. Next note that for $m = 0$, Case 1 and Case 2 are the same.

Now consider the case $m = 1$. Estimation points r_j are the solutions of

$${}_2F_1\left(-2,\frac{k}{2}+2;1;1-\frac{r^2}{h^2}\right)=0, \tag{3.53}$$

which are given by

$$r_j^2=\frac{\pm 2\sqrt{2}\sqrt{(k+4)(k+2)}+(k+4)(k+2)}{(k+6)(k+4)}h^2,\quad j=1,2. \tag{3.54}$$

The weights c_j are then given by

$$c_j=\frac{2^{k/2-1}}{\left(\frac{k}{2}+2\right)^2\frac{r_j^2}{h^2}\left(1-\frac{r_j^2}{h^2}\right)2^2\left[{}_2F_1\left(-1,\frac{k}{2}+3;2;1-\frac{r_j^2}{h^2}\right)\right]^2}, \tag{3.55}$$

where ${}_2F_1\left(-1,\frac{k}{2}+3;2;1-\frac{r^2}{h^2}\right)=\frac{1}{4h^2}\left(-h^2(k+2)+r_j^2(k+6)\right)$.

Thus, it follows that

$$\begin{aligned} c_j &= \frac{2^{k/2+1}h^6}{\left(\frac{k}{2}+2\right)^2 r_j^2\left(1-\frac{r_j^2}{h^2}\right)\left(r_j^2(k+6)-h^2(k+2)\right)^2} \\ &= \frac{h^6}{9r_j^2\left(1-\frac{r_j^2}{h^2}\right)(2r_j^2-h^2)^2}. \end{aligned} \tag{3.56}$$

For the case $m = 2$ $(k = 2)$, r_j, the solutions of ${}_2F_1\left(-3, \frac{k}{2}+3; 1; 1-\frac{r_j^2}{h^2}\right) = 0$, are

$$\begin{aligned} r_1^2 &= \frac{1}{2}h^2, \\ r_2^2 &= \frac{5+\sqrt{15}}{10}h^2, \\ r_3^2 &= \frac{5-\sqrt{15}}{10}h^2. \end{aligned} \tag{3.57}$$

The weights c_j are then given by

$$c_j = \frac{2^{k/2-1}}{\left(\frac{k}{2}+3\right)^2 \frac{r_j^2}{h^2}\left(1-\frac{r_j^2}{h^2}\right)\left[3\,{}_2F_1\left(-2, \frac{k}{2}+4; 2; 1-\frac{r_j^2}{h^2}\right)\right]^2}, \tag{3.58}$$

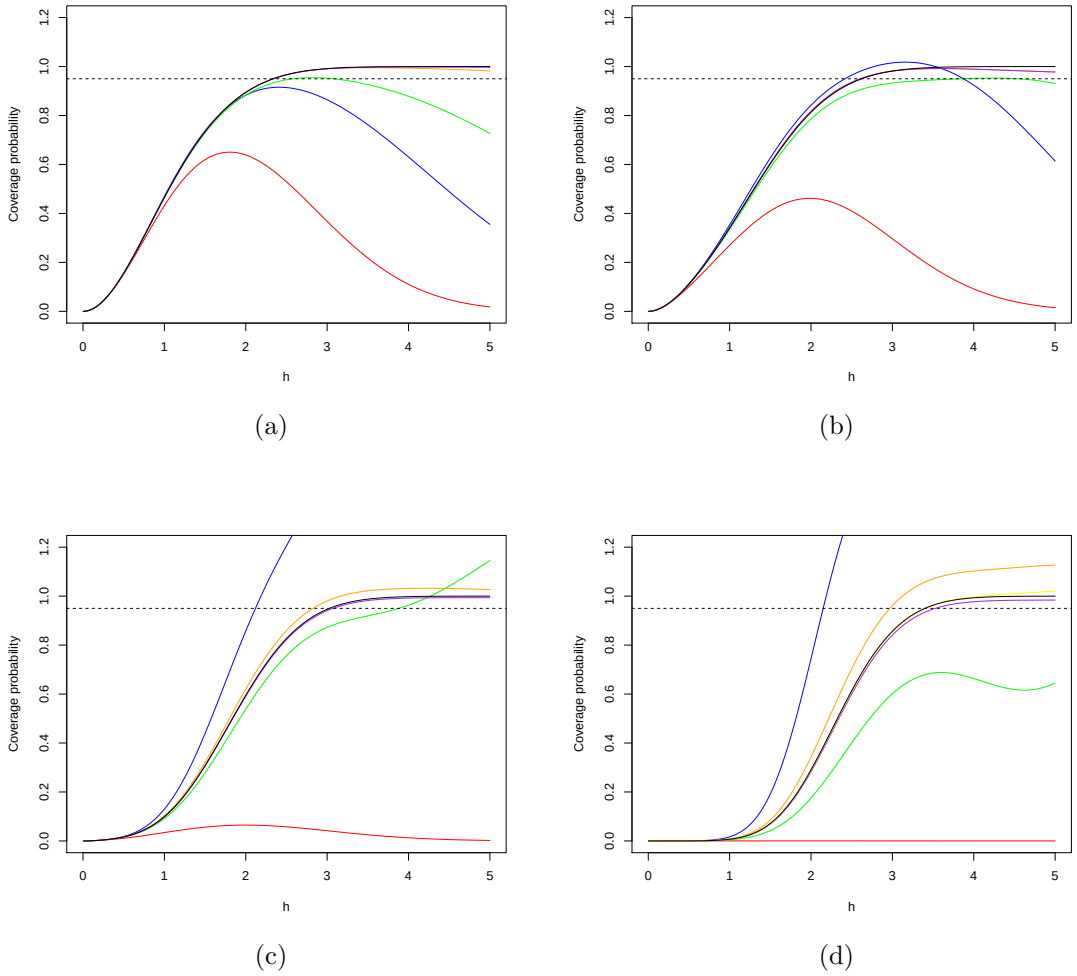
where ${}_2F_1\left(-2, \frac{k}{2}+4; 2; 1-\frac{r_j^2}{h^2}\right) = \frac{1}{h^4}\left(h^4 - 5h^2r_j^2 + 5r_j^4\right)$. Thus,

$$c_j = \frac{h^{10}}{16 \cdot 9r_j^2\left(1-\frac{r_j^2}{h^2}\right)(h^4 - 5h^2r_j^2 + 5r_j^4)^2}. \tag{3.59}$$

Lastly for $m = 3$ we have the following values

$$\begin{aligned} r_{1,2}^2 &= \frac{35+\sqrt{35(15\pm 2\sqrt{30})}}{70}h^2, \\ r_{3,4}^2 &= \frac{35-\sqrt{35(15\pm 2\sqrt{30})}}{70}h^2, \end{aligned} \tag{3.60}$$

$$\begin{aligned} c_j &= \frac{2^{k/2-1}}{\left(4+\frac{k}{2}\right)^2 \frac{r_j^2}{h^2}\left(1-\frac{r_j^2}{h^2}\right)\left[4\,{}_2F_1\left(-3, 6; 2; 1-\frac{r_j^2}{h^2}\right)\right]^2} \\ &= \frac{h^{14}}{2^4 \cdot 5^2 r_j^2\left(1-\frac{r_j^2}{h^2}\right)\left(14r_j^6 - 21h^2.r_j^4 + 9h^4r_j^2 - h^6\right)^2}. \end{aligned} \tag{3.61}$$



(a) (b)

(c) (d)

Figure 3.4: Approximation for coverage probability for different precision m. (a) $p = 3$, (b) $p = 5$, (c) $p = 10$, (d) $p = 20$ and $k = 2$. Red line corresponds to $m = 0$, blue $m = 1$, green $m = 2$, orange $m = 5$, purple $m = 10$, yellow $m = 20$, black - exact coverage probability.

From the plots in the Figure 3.4 it seems that for all considered p approximation with $m = 10$ performs well. We use approximation with level of precision $K = 4m + 3$, where $m = 10$ in the Tables 3.1-3.4. Approximation (3.43) gives values of h that are very close to the exact value of h obtained by equating (3.46) to $1 - \alpha$, especially for small p. For all the performed simulations h in (3.43) was significantly smaller than h as a solution of (3.9) while still maintaining $1-\alpha$ coverage probability.

Table 3.1: Estimated coverage probabilities for the case $p = 3$, $k = 2$, $m = 10$.

$\boldsymbol{\theta}$	$h_{\text{exact}} = 2.34$	$h_{\text{app}} = 2.35$	$h_{\text{crude}} = 3.16$	$h_{\text{naive}} = 2.45$
$(0, 0, 0)$	0.957	0.958	0.997	0.973
$(0, 0.25, 0.5)$	0.946	0.947	0.992	0.959
$(0, 5, 10)$	0.950	0.952	0.994	0.96
$(0, 0, 2)$	0.936	0.937	0.990	0.952
$(0, 0, 5)$	0.932	0.934	0.994	0.948

Table 3.2: Estimated coverage probabilities for the case $p = 5$, $k = 2$, $m = 10$.

$\boldsymbol{\theta}$	$h_{\text{exact}} = 2.63$	$h_{\text{app}} = 2.64$	$h_{\text{crude}} = 3.37$	$h_{\text{naive}} = 2.45$
$(0, 0, 0, 0, 0)$	0.947	0.947	0.991	0.909
$(0, 0.25, 0.5, 0.7, 1)$	0.965	0.966	0.999	0.938
$(1, 3, 5, 7, 9)$	0.968	0.969	0.996	0.948
$(0, 0, 2, 2, 2)$	0.965	0.968	0.997	0.947
$(0, 0, 5, 5, 5)$	0.984	0.984	0.997	0.967

Table 3.3: Estimated coverage probabilities for the case $p = 10$, $k = 2$, $m = 10$.

$\boldsymbol{\theta}$	$h_{\text{exact}} = 3.03$	$h_{\text{app}} = 3.08$	$h_{\text{crude}} = 3.68$	$h_{\text{naive}} = 2.45$
$(0, 0, 0, 0, 0, 0, 0, 0, 0, 0)$	0.941	0.950	0.993	0.817
$(0, 0, 0, 0, 0, 1, 1, 1, 1, 1)$	0.956	0.963	0.990	0.844
$(0, 0, 0, 0, 0, 5, 5, 5, 5, 5)$	0.984	0.988	0.999	0.929
$(0.1, 0.2, \ldots, 1)$	0.952	0.958	0.988	0.820
$(1, 2, \ldots, 10)$	0.992	0.995	0.998	0.943

Table 3.4: Estimated coverage probabilities for the case $p = 20$, $k = 2$, $m = 10$.

$\boldsymbol{\theta}$	$h_{\text{exact}} = 3.4$	$h_{\text{app}} = 3.53$	$h_{\text{crude}} = 3.98$	$h_{\text{naive}} = 2.45$
$(0, \ldots, 0)$	0.953	0.972	0.995	0.572
$(0, \ldots, 0, 1, \ldots, 1)$	0.952	0.963	0.985	0.691
$(0, \ldots, 0, 5, \ldots, 5)$	0.979	0.985	0.999	0.836
$(0.1, 0.2, \ldots, 2)$	0.957	0.971	0.995	0.695
$(1, 2, \ldots, 20)$	0.996	0.998	0.999	0.945

Heuristically, can we choose m to provide a good approximation in the following way. We can calculate the approximation for a range of m, plot them, and choose

the one that converges to a normal cdf. Then the radius of our confidence set, h, for this m-approximation is the solution of (3.43).

Case general p and $k \geq 3$

We will now address the general case. Without loss of generality assume that selected set is $I_j = \{1, 2, \ldots, k\}$ and not selected set $I_j^c = \{k+1, \ldots, p\}$.

$$P\left(\sum_{i=1}^{k}(X_{(i)} - \theta_{(i)})^2 \leq h^2\right) \geq \binom{p}{k}\left(\frac{1}{2\pi}\right)^{(k-1)/2} \int_0^h r^{k-1}\phi(r) \int_0^\pi \sin^{k-2}\varphi_{k-1} \ldots$$
$$\times \int_0^\pi \sin\varphi_2 \int_0^{2\pi} \Phi^{p-k}\left(r \min\left\{\sin\varphi_1 \sin\varphi_2 \ldots \sin\varphi_{k-1},\right.\right.$$
$$\left.\left.\{\cos\varphi_{l-1}\sin\varphi_l \ldots \sin\varphi_{k-1}\}_{l=2}^{k}\right\}\right)\, d\varphi_1 \ldots\, d\varphi_{k-1}\, dr.$$

First consider the case $k = 3$. Then

$$P\left(\sum_{i=1}^{3}(X_{(i)} - \theta_{(i)})^2 \leq h^2\right) \geq \binom{p}{k}\left(\frac{1}{2\pi}\right)^{(k-1)/2} \int_0^h r^{k-1}\phi(r) \int_0^\pi \sin\varphi_2$$
$$\times \int_0^{2\pi} \Phi^{p-k}\left(r \min\left\{\sin\varphi_1 \sin\varphi_2; \cos\varphi_1 \sin\varphi_2; \cos\varphi_2\right\}\right)\, d\varphi_1\, d\varphi_2\, dr.$$

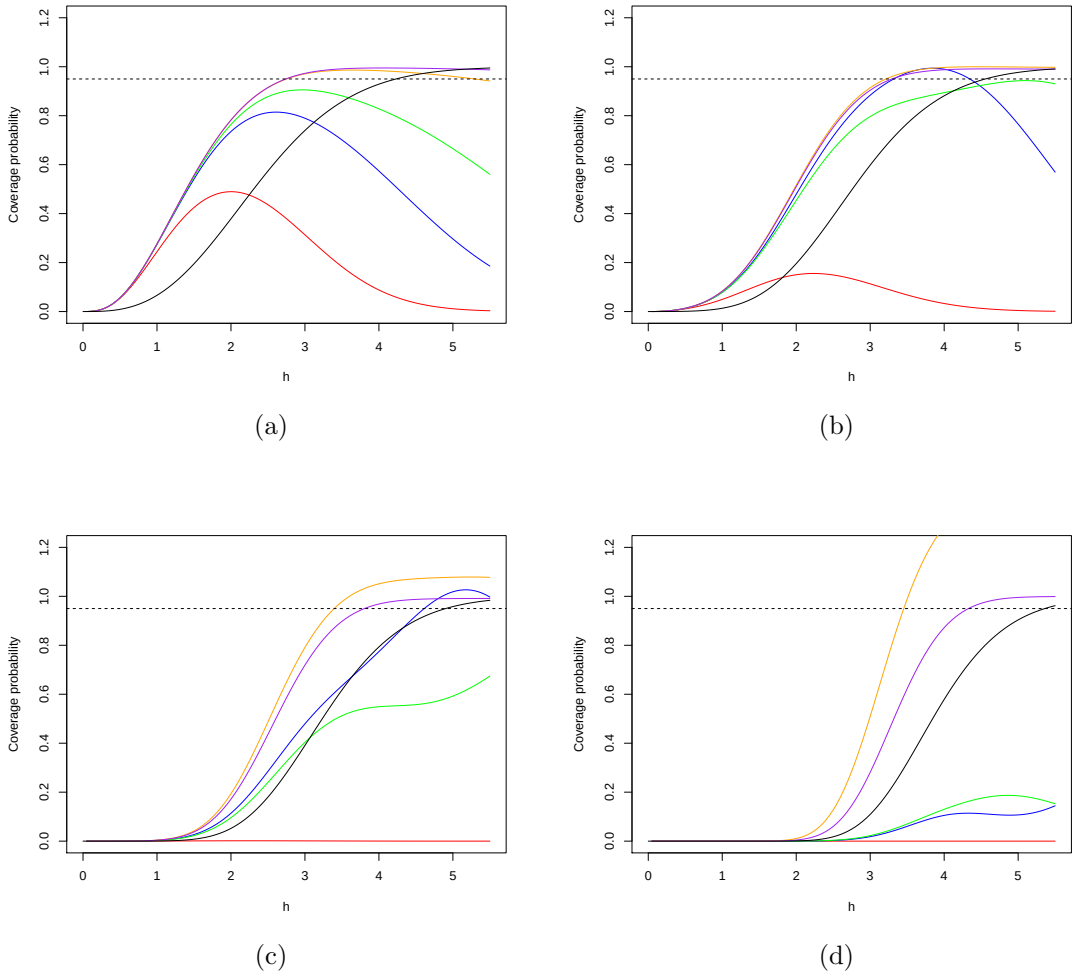
Even in this simple case for selecting only $k = 3$ means, it will be hard to split the integral with respect to φ_1 into intervals corresponding to the minimum of $\{\sin\varphi_1 \sin\varphi_2; \cos\varphi_1 \sin\varphi_2; \cos\varphi_2\}$.

Approximation (3.2.1) for the coverage probability becomes

$$P\left(\sum_{i=1}^{k}(X_{(i)} - \theta_{(i)})^2 \leq h^2\right) \geq \binom{p}{k}\left(\frac{1}{2\pi}\right)^{(k-1)/2}\left(\frac{h}{\sqrt{2}}\right)^k \sum_{g=1}^{4(m+1)} \frac{\pi}{2(m+1)} \sum_{j=1}^{m+1} c_j\phi(r_j)$$
$$\times \sum_{i}^{2m+2} b_{2i}\Phi^{p-k}\left(r_j \min\left\{\cos\frac{\pi g}{2(m+1)}\sin\varphi_{2,i}; \sin\frac{\pi g}{2(m+1)}\sin\varphi_{2,i}; \cos\varphi_{2,i}\right\}\right), \tag{3.62}$$

where r_j's are given in (3.50), (3.54), and (3.57) and c_j's are given in (3.51), (3.56), and (3.58) for different m.

In Figures 3.5 and 3.6 approximation (3.62) is compared to approximation (3.9). For Figures 3.5 (a)-(b) approximation (3.62) for $m = 8$ performs quite well. For 3.5 (c)-(d) it is not clear if approximation for $m = 8$ is good enough, but we still use it in Tables 3.5 - 3.7. There are numerical issues with calculating the approximation (3.62) for high m, since zeros of (3.25) are numerically evaluated as complex numbers even though it is known that zeroes of Jacobi polynomials $Q_p^{\alpha,\beta}$ for $\alpha,\ \beta > -1$ are real and belong to $[-1, 1]$. To evaluate approximation (3.62) for high m a numerically stable way to evaluate the zeroes or (3.25) is needed.



(a) (b)

(c) (d)

Figure 3.5: Approximation for coverage probability for different precision m. (a) $p = 5$, (b) $p = 10$, (c) $p = 20$, (d) $p = 50$ and $k = 3$. Red line corresponds to $m = 0$, blue $m = 1$, green $m = 2$, orange $m = 5$, purple $m = 8$, black - crude approximation (3.9) of coverage probability.

For the case $p = 5$ and $k = 3$ (Table 3.5) approximation (3.62) provides similar results to the naive set. As we increase p, especially for $p = 50$, the naive set performs poorly, which is expected, while the approximated set maintains coverage probability of $1 - \alpha$ and is much smaller than the approximated set (3.9) which is very conservative yielding coverage probabilities close to one.

Table 3.5: Estimated coverage probabilities for the case $p = 5$, $k = 3$, $m = 8$.

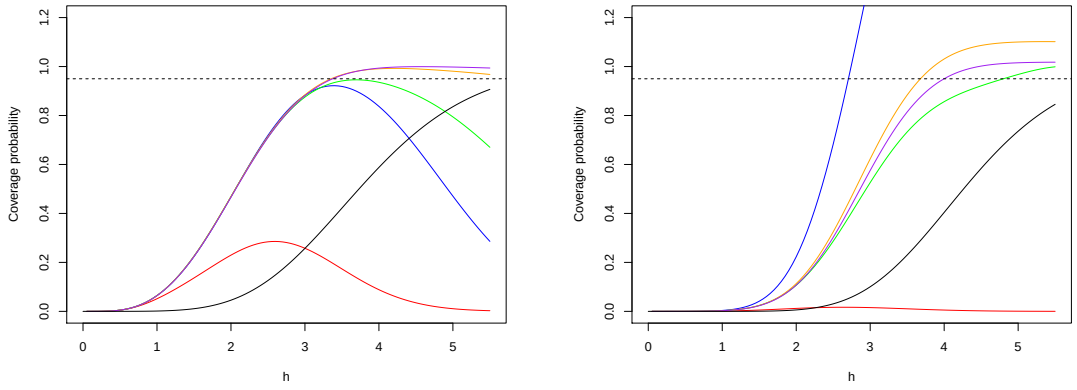
$\boldsymbol{\theta}$	$h_{\text{app}} = 2.75$	$h_{\text{crude}} = 4.23$	$h_{\text{naive}} = 2.8$
$(0, 0, 0, 0, 0)$	0.963	0.999	0.968
$(0, 0.25, 0.5, 0.75, 1)$	0.963	0.999	0.971
$(1, 3, 5, 7, 9)$	0.948	1	0.953
$(0, 0, 0, 2, 2)$	0.954	1	0.960
$(0, 0, 0, 5, 5)$	0.941	0.999	0.945

Table 3.6: Estimated coverage probabilities for the case $p = 10$, $k = 3$, $m = 8$.

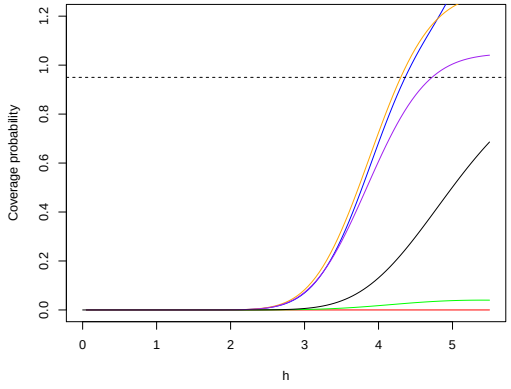
$\boldsymbol{\theta}$	$h_{\text{app}} = 3.3$	$h_{\text{crude}} = 4.56$	$h_{\text{naive}} = 2.8$
$(0, 0, 0, 0, 0, 0, 0, 0, 0, 0)$	0.962	1	0.850
$(0, 0, 0, 0, 0, 1, 1, 1, 1, 1)$	0.958	0.998	0.878
$(0, 0, 0, 0, 0, 5, 5, 5, 5, 5)$	0.994	1	0.965
$(0.1, 0.2, \ldots, 1)$	0.959	1	0.872
$(1, 2, \ldots, 10)$	0.989	0.999	0.946

Table 3.7: Estimated coverage probabilities for the case $p = 50$, $k = 3$, $m = 8$.

$\boldsymbol{\theta}$	$h_{\text{app}} = 4.35$	$h_{\text{crude}} = 5.36$	$h_{\text{naive}} = 2.8$
$(0, \ldots, 0)$	0.948	1	0.151
$(0, \ldots, 0, 1, \ldots, 1)$	0.974	0.999	0.355
$(0, \ldots, 0, 5, \ldots, 5)$	0.992	0.999	0.486
$(0.1, 0.2, \ldots, 2)$	0.994	1	0.706
$(1, 2, \ldots, 20)$	0.999	1	0.949



(a) (b)



(c)

Figure 3.6: Approximation for coverage probability for different precision m. (a) $p = 10$, (b) $p = 20$, (c) $p = 50$, and $k = 5$. Red line corresponds to $m = 0$, blue $m = 1$, green $m = 2$, orange $m = 5$, purple $m = 8$, black - crude approximation (3.9) of coverage probability.

Table 3.8: Estimated coverage probabilities for the case $p = 10$, $k = 5$, $m = 8$.

$\boldsymbol{\theta}$	$h_{\text{app}} = 3.4$	$h_{\text{crude}} = 5.6$	$h_{\text{naive}} = 3.33$
$(0, 0, 0, 0, 0, 0, 0, 0, 0, 0)$	0.952	1	0.945
$(0, 0, 0, 0, 0, 1, 1, 1, 1, 1)$	0.946	1	0.939
$(0, 0, 0, 0, 0, 5, 5, 5, 5, 5)$	0.956	1	0.948
$(0.1, 0.2, \ldots, 1)$	0.939	1	0.928
$(1, 2, \ldots, 10)$	0.948	1	0.941

Table 3.9: Estimated coverage probabilities for the case $p = 20$, $k = 5$, $m = 8$.

$\boldsymbol{\theta}$	$h_{\text{app}} = 4$	$h_{\text{crude}} = 6.39$	$h_{\text{naive}} = 3.33$
$(0, \ldots, 0)$	0.942	1	0.730
$(0, \ldots, 0, 1, \ldots, 1)$	0.948	1	0.781
$(0, \ldots, 0, 5, \ldots, 5)$	0.985	1	0.939
$(0.1, 0.2, \ldots, 2)$	0.954	1	0.817
$(1, 2, \ldots, 20)$	0.999	1	0.949

Table 3.10: Estimated coverage probabilities for the case $p = 50$, $k = 5$, $m = 8$.

$\boldsymbol{\theta}$	$h_{\text{app}} = 4.75$	$h_{\text{crude}} = 6.94$	$h_{\text{naive}} = 3.33$
$(0, \ldots, 0)$	0.918	1	0.172
$(0, \ldots, 0, 1, \ldots, 1)$	0.954	1	0.363
$(0, \ldots, 0, 5, \ldots, 5)$	0.984	1	0.612
$(0.1, 0.2, \ldots, 2)$	0.991	1	0.753
$(1, 2, \ldots, 20)$	1	1	0.945

The following table, which provides coefficients b_i (3.15) and evaluation points y_i, solutions of (3.17), is taken from Mustard (1964).

Table 3.11: Coordinates and coefficients for the k-dimensional sphere angular integration rule $\int_0^\pi \sin^\nu \phi g(\phi)\, d\varphi \approx \sum_{i=1}^{2m+2} b_i g(\varphi_i)$, $\nu = 1, 2, \ldots, k-2$, where $y = \cos\varphi$. Values in the table are given for the case $m = 0$.

ν	Polynomial $Q_{2m+2}^{(\nu-1)/2}$	Coordinates y_i	Coordinates b_i
$m = 0$			
1	$\frac{3}{2}\left(y^2 - \frac{1}{3}\right)$	$\pm\frac{1}{\sqrt{3}}$	1
2	$\frac{5}{2}\left(y^2 - \frac{1}{4}\right)$	$\cos\frac{\pi}{3}, \cos\frac{2\pi}{3}$	$\frac{\pi}{4}$
3	$\frac{15}{4}\left(y^2 - \frac{1}{5}\right)$	$\pm\frac{1}{\sqrt{5}}$	$\frac{2}{3}$
4	$\frac{21}{4}\left(y^2 - \frac{1}{6}\right)$	$\pm\frac{1}{\sqrt{6}}$	$\frac{3\pi}{16}$
5	$7\left(y^2 - \frac{1}{7}\right)$	$\pm\frac{1}{\sqrt{7}}$	$\frac{8}{15}$
5	$9\left(y^2 - \frac{1}{8}\right)$	$\pm\frac{1}{\sqrt{8}}$	$\frac{5\pi}{32}$
$m = 1$			
1	$\frac{35}{8}\left(y^4 - \frac{6}{7}y^2 + \frac{3}{35}\right)$	$\pm\sqrt{\frac{15+2\sqrt{30}}{35}}$	$\frac{49}{6(18+\sqrt{30})}$
		$\pm\sqrt{\frac{15-2\sqrt{30}}{35}}$	$\frac{49}{6(18-\sqrt{30})}$
2	$\frac{63}{8}\left(y^4 - \frac{3}{4}y^2 + \frac{1}{16}\right)$	$\cos\frac{\pi}{5}, \cos\frac{4\pi}{5}$	$\frac{\pi}{8}\left(1 - \frac{1}{\sqrt{5}}\right)$
		$\cos\frac{2\pi}{5}, \cos\frac{3\pi}{5}$	$\frac{\pi}{8}\left(1 + \frac{1}{\sqrt{5}}\right)$
3	$\frac{105}{8}\left(y^4 - \frac{2}{3}y^2 + \frac{1}{21}\right)$	$\pm\sqrt{\frac{7+\sqrt{28}}{21}}$	$\frac{6}{5}\left(\frac{1}{5-\sqrt{7}}\right)$
		$\pm\sqrt{\frac{7-\sqrt{28}}{21}}$	$\frac{6}{5}\left(\frac{1}{5+\sqrt{7}}\right)$
4	$\frac{165}{8}\left(y^4 - \frac{3}{5}y^2 + \frac{3}{80}\right)$	$\pm\sqrt{\frac{6+\sqrt{21}}{20}}$	$\frac{125\pi}{32(63+8\sqrt{21})}$
		$\pm\sqrt{\frac{6-\sqrt{21}}{20}}$	$\frac{125\pi}{32(63-8\sqrt{21})}$
5	$\frac{495}{16}\left(y^4 - \frac{6}{11}y^2 + \frac{1}{33}\right)$	$\pm\sqrt{\frac{9+4\sqrt{3}}{33}}$	$\frac{242}{105(14+5\sqrt{3})}$
		$\pm\sqrt{\frac{9-4\sqrt{3}}{33}}$	$\frac{242}{105(14-5\sqrt{3})}$
6	$\frac{715}{16}\left(y^4 - \frac{1}{2}y^2 + \frac{1}{40}\right)$	$\pm\sqrt{\frac{5+\sqrt{15}}{20}}$	$\frac{35\pi}{128(6+\sqrt{15})}$
		$\pm\sqrt{\frac{5-\sqrt{15}}{20}}$	$\frac{35\pi}{128(6-\sqrt{15})}$

CHAPTER 4

CONFIDENCE SETS

This chapter is devoted to confidence sets for estimating the selected parameters. Throughout this chapter we will assume $\mathbf{X} \sim N_p(\boldsymbol{\theta}, \sigma^2\mathbf{I})$, where $\sigma^2 = 1$ unless otherwise specified. Most of the results are obtained for comparisons to the naive set (3.3) for estimating $\theta_{(1)} = \sum_{i=1}^{p} \theta_i I(X_i = X_{(1)})$ where $X_{(1)} \geq \cdots \geq X_{(p)}$.

4.1 Minimaxity

Standard approach of decision theory uses a loss function suited to a particular estimation problem and examines expected loss, that is, the risk of a statistical procedure. In this section we undertake this approach to set estimation under the normal model for selection. The advantage of this is that it provides a simple way to address optimality properties such as minimaxity and admissibility, also under some specified loss it is easy to obtain Bayes and generalized Bayes rules.

4.1.1 The loss function for set estimation

Unfortunately, there is no well established loss function for working with confidence sets. Most researchers either use componentwise loss or a linear combination loss function which combines some measure of volume of the set, $\nu(C)$, and coverage probability or indicator of inclusion, $I(\theta \in C)$, which has the following form

$$L(\theta, C) = b\nu(C) - I(\theta \in C), \tag{4.1}$$

where b is a fixed constant. The second loss is more popular but it gets roundly criticized a lot. For example, Casella et al. (1993) elaborated on the paradox

pointed out by Brown that shows that under this loss empty set dominates t-interval when sample variance is large. They also suggest using a modified loss, $L(\theta, C) = S\left[\nu(C)\right] - I\left(\theta \in C\right)$, where $S(\cdot)$ is a size function, which avoids the above paradox for Student's t-interval for some S. The following size function $S(\cdot)$ avoids the paradox

$$S_a(\nu) = \frac{\nu}{a+\nu},$$

where a is some positive number.

The loss in (4.1) was used by Joshi (1969), who was working on minimaxity and admissibility of the usual confidence set; Winkler (1972) discusses interval estimation in decision-theoretic framework under this loss; Cohen & Strawderman (1973) provide sufficient conditions for admissibility of the best invariant confidence interval under a set of losses containing (4.1), and Meeden & Vardeman (1985) who explored the connection between Bayes rules under (4.1) and admissibility.

Casella & Hwang (1991) considered the connection between the linear combination loss function and component loss problem. They identified procedures which perform well against the linear combination loss function and also perform well against each measure for a particular linear combination of components of the loss. They were mostly interested in exploring the equivalence of the loss functions with respect to minimaxity. They were able to derive necessary and sufficient conditions for the set to be minimax under both losses.

4.1.2 Minimaxity for a confidence procedure

There are two major approaches for defining minimaxity in the context of confidence sets estimation. The first way defines minimaxity as a part of the component loss problem, while the other way attaches the minimaxity to a given loss (this is the

common definition for minimaxity of the estimate). Specifically,

Definition 1. *A confidence procedure C^* is said to be α-minimax under the loss $L(\theta, C)$ if, for a given confidence level α,*

$$P_\theta(C^*) \geq 1 - \alpha, \text{ for all } \theta,$$

$$\sup_\theta E_\theta \left(b\nu(C^*)\right) = \inf_\phi \sup_\theta E_\theta \left(b\nu(C)\right),$$

where b is some constant.

For the second definition we consider loss (4.1), the common loss to consider when dealing with estimating confidence sets. A more typical definition of minimaxity is the following.

Definition 2. *A confidence procedure C^* is said to be k-minimax under the loss $L(\theta, C)$ if*

$$\sup_\theta EL(\theta, C^*) = \inf_\phi \sup_\theta EL(\theta, C).$$

In the following we will talk about minimaxity under both definitions given above. Definition 1 is somewhat stronger, as it requires the coverage probability to be at least $1 - \alpha$. Which makes sense since we want our set to have nominal coverage. Definition 2 is not as strong in this context. Minimaxity in this case depends on the loss you are considering and there is no restriction on coverage. We will see that even though the usual confidence set will not be minimax under the Definition 1, it will be minimax under less strict Definiton 2.

Minimaxity for the case $p = 2$ and $k = 1$

By definition, for a set to be α-minimax (see Casella & Hwang (1991)), it needs to maintain a coverage probability $\geq 1-\alpha$ and minimize the "worst" (largest) expected

volume. Let's consider the sets (intervals) of the form $a \leq X_{(1)} - \theta_{(1)} \leq b$, where a and b are some non-negative constants. To find a minimax rule, we need to minimize $a + b$, the length of this interval, given that

$$P\left(a \leq X_{(1)} - \theta_{(1)} \leq b\right) \geq 1 - \alpha,$$

or similarly $P(a \leq X_{(1)} - \theta_{(1)} \leq b) = \Phi^2(b) - \Phi^2(a) = (\Phi(b) - \Phi(a))(\Phi(b) + \Phi(a)) \geq 1 - \alpha$. To minimize $a + b$ will have to have $a = -b$. Thus $\Phi(b) - \Phi(-b) \geq 1 - \alpha$, and so $b = \Phi^{-1}(\alpha/2)$, which corresponds to the naive set $|X_{(1)} - \theta_{(1)}| \leq \Phi^{-1}(\alpha/2)$. Thus, naive set is minimax for $p = 2$ and $k = 1$.

Minimaxity for general p and k

The usual confidence set for estimating $\boldsymbol{\theta}_s$ is

$$\phi_0 = \begin{cases} 1, & \text{if } \sum_{i=1}^{p} \sum_{j=1}^{k} (X_i - \theta_i)^2 \, I\left(X_i = X_{(j)}\right) \leq h^2 \\ 0, & \text{otherwise.} \end{cases}$$

Consider the loss (4.1) rewritten in the following way

$$\begin{aligned} L(\boldsymbol{\theta}_s, \phi) &= b\upsilon(\boldsymbol{\theta}_s : \boldsymbol{\theta}_s \in C) - I\left(\boldsymbol{\theta}_s \in C\right) \\ &= b\upsilon(\phi) - \phi, \end{aligned}$$

where $\upsilon(\phi) = \int \phi(x, \theta)\, d\theta$ is the volume of the confidence interval (in our case it is the length of the interval), and b is some constant.

We want to find the Bayes rule under this loss. Consider independent priors for θ_i's: $\xi(\theta_i) = \frac{1}{\sqrt{2\pi\tau^2}} \exp\left\{-\frac{\theta_i^2}{2\tau^2}\right\}$, $i = 1, \ldots, p$. To find the Bayes rule, we need to

minimize Bayes risk with respect to $\boldsymbol{\theta}_s$:

$$\begin{aligned}
r(\xi, \boldsymbol{\theta}) &= \int_{\mathcal{X}} \int_{\Theta} L(\boldsymbol{\theta}_s, \phi) \pi(\boldsymbol{\theta}|\mathbf{x}) m(\mathbf{x})\, d\boldsymbol{\theta}\, d\mathbf{x} \\
&= \int_{\mathcal{X}} \int_{\Theta} \left[b\nu(\boldsymbol{\theta}_s \in C) - I(\boldsymbol{\theta}_s \in C) \right] \pi(\boldsymbol{\theta}|\mathbf{x}) m(\mathbf{x})\, d\mathbf{x} \\
&= \int_{\mathcal{X}} \left[\int_{\Theta} b\nu(\boldsymbol{\theta}_s \in C) \pi(\boldsymbol{\theta}|\mathbf{x})\, d\boldsymbol{\theta} - \int_{\Theta} I(\boldsymbol{\theta} \in C) \pi(\boldsymbol{\theta}|\mathbf{x})\, d\boldsymbol{\theta} \right] m(\mathbf{x})\, d\mathbf{x} \\
&= \int_{\mathcal{X}} \left[b\nu(\boldsymbol{\theta}_s \in C) - \int_{\Theta} I(\boldsymbol{\theta} \in C) \pi(\boldsymbol{\theta}|\mathbf{x})\, d\boldsymbol{\theta} \right] m(\mathbf{x})\, d\mathbf{x} \\
&= \int_{\mathcal{X}} \left[b \int_{\Theta} I(\boldsymbol{\theta}_s \in C)\, d\boldsymbol{\theta} - \int_{\Theta} I(\boldsymbol{\theta} \in C) \pi(\boldsymbol{\theta}|\mathbf{x})\, d\boldsymbol{\theta} \right] m(\mathbf{x})\, d\mathbf{x} \\
&= \int_{\mathcal{X}} \left[\int_{\Theta} \left\{ b - \pi(\boldsymbol{\theta}|\mathbf{x}) \right\} I(\boldsymbol{\theta}_s \in C)\, d\boldsymbol{\theta} \right] m(\mathbf{x})\, d\mathbf{x} \\
&= \int_{\mathcal{X}} \left[\int_{\Theta} \left\{ b \prod_{j=1}^{k} I(\theta_{(j)} \in C) - \prod_{i=1}^{p} \pi(\theta_i|x) \prod_{j=1}^{k} I(\theta_{(j)} \in C) \right\} d\boldsymbol{\theta} \right] m(\mathbf{x})\, d\mathbf{x} \\
&= \int_{\mathcal{X}} \left[\int_{\Theta} \left\{ b \prod_{j=1}^{k} I(\theta_{(j)} \in C) \right\} d\boldsymbol{\theta}_s - \int_{\Theta} \prod_{j=1}^{k} \left\{ \pi(\theta_{(j)}|x) I(\theta_{(j)} \in C) \right\} d\boldsymbol{\theta}_s \right] m(\mathbf{x})\, d\mathbf{x} \\
&= \int_{\mathcal{X}} \left[\int_{\Theta} b\phi\, d\boldsymbol{\theta} - \int_{\Theta} \prod_{j=1}^{k} \pi(\theta_{(j)}|x) \phi\, d\boldsymbol{\theta} \right] m(\mathbf{x})\, d\mathbf{x} \\
&= \int_{\mathcal{X}} \int_{\Theta} \left[b - \prod_{j=1}^{k} \pi(\theta_{(j)}|x) \right] \phi\, d\boldsymbol{\theta} m(\mathbf{x})\, d\mathbf{x},
\end{aligned}$$

where $\phi = \prod_{j=1}^{k} I\left(\theta_{(j)} \in C\right)$.

Since ϕ is a decision rule (i.e. only takes values 0 and 1), the Bayes risk is minimized at

$$\phi_\tau = \begin{cases} 1, & \text{if } b \leq \prod_{j=1}^{k} \pi(\theta_{(j)}|x) \\ 0, & \text{otherwise.} \end{cases} \tag{4.2}$$

Thus, the Bayes rule is given by

$$\phi_\tau = \begin{cases} 1, & \text{if } \sum_{i=1}^{p} \sum_{j=1}^{k} \left(\theta_i - (1-B)X_i\right)^2 I(X_i = X_{(j)}) \leq h^{*2}, \\ 0, & \text{otherwise,} \end{cases} \tag{4.3}$$

where $h^{*2} = 2(1-B) \log \left[b \left(2\pi(1-B)\right)^{k/2} \right]^{-1}$.

For minimaxity we need

$$\lim_{\tau^2\to\infty} E\nu(C^\pi) \geq E\nu(C_0) = \frac{\pi^{k/2}}{\Gamma(\frac{k}{2}+1)}h^k, \tag{4.4}$$

$$\lim_{\tau^2\to\infty} EI(\theta_s \in C^\pi) \geq EI(\theta_s \in C_0). \tag{4.5}$$

Consider the following limit

$$\begin{aligned}
\lim_{\tau^2\to\infty} E\nu(C^\pi) &= \lim\left[\frac{\pi^{k/2}h^{*k}}{\Gamma\left(\frac{k}{2}+1\right)}\right] = \frac{\pi^{k/2}}{\Gamma\left(\frac{k}{2}+1\right)}\lim h^{*k} \geq \frac{\pi^{k/2}}{\Gamma\left(\frac{k}{2}+1\right)}h^k, \\
&\Rightarrow h^{*k} \geq h^k, \\
&\Rightarrow \left(2\log\left[b(2\pi)^{k/2}\right]^{-1}\right)^{k/2} \geq h^k, \\
&\Rightarrow b \leq \left(\frac{1}{2\pi}\right)^{k/2} e^{-h^2/2}.
\end{aligned} \tag{4.6}$$

Furthermore notice that

$$EI(\boldsymbol{\theta}_s \in C^\pi) = P(\boldsymbol{\theta}_s \in C^\pi) = \sum_{j=1}^{\binom{p}{k}} \int_{z_1^2+\cdots+z_k^2\leq h^{*2}} \cdots \int \prod_{m\in I_j^C} \Phi\left(\min_{l\in I_j}\{z_l+\theta_l\} - \theta_m\right) \prod_{l\in I_j} \phi(z_l)\, dz_l$$

$$\begin{aligned}
&\geq EI(\boldsymbol{\theta}_s \in C_0) \\
&\Rightarrow h^{*2} \geq h^2.
\end{aligned}$$

Theorem 5. *For selecting any k populations out of p, the usual set C_0 is minimax under the loss $L(\boldsymbol{\theta}_s, C) = b\nu(C) - I(\boldsymbol{\theta}_s \in C)$, where $b = \left(\frac{1}{2\pi}\right)^{k/2} e^{-h^2/2}$ (under Definition 2).*

If now instead we consider minimaxity under stronger Definition 1, we will see that the usual confidence set maintains $1-\alpha$ coverage probability only for $p = 2$. Thus, it turns out that the set will be minimax in this case (see Section 4.2). An alternative proof of minimaxity is given in the next section.

Under Definition 1, for selecting $k = 1$ populations out of $p = 2$, the usual set C_0 is minimax.

Note that in Casella & Hwang (1991) it is shown that the usual confidence set for non-selection problem is minimax if $b = \left(\frac{1}{2\pi}\right)^{p/2} e^{-h^2/2}$, which is exactly what we have for selection problem with p replaced by k. This suggests that after conditioning on selected populations, we are only working with selected k populations, and selection as it is does not play a role anymore.

4.2 Best equivariant rule is minimax

The rule ϕ_0 and the loss $L(\theta, \phi_0)$ are invariant under the location-scale group of transformations, that is,

$$(\theta, \sigma) \to (c\theta + \sigma, c\sigma).$$

Here we consider σ is unknown.

The right Haar invariant measure for the location-scale group is (Lehmann & Casella (2003))

$$\pi^r(\theta, \sigma) = \frac{1}{\sigma}.$$

To find best equivariant rule, we need to find the Bayes rule with respect to this right Haar invariant density (see Hunt-Stein Theorem 5.47 in Liese & Miescke (2008))

$$\begin{aligned}
\int\int\int L(\theta, \phi) p(x, \theta) \pi^r(\theta, \sigma)\, d\theta\, d\sigma\, dx = \\
&= \int\int \sum_{i=1}^{p} \int \left\{ \frac{b}{\sigma} Vol(\phi) - \phi \right\} I(x_i = x_{(1)}) p(x|\theta, \sigma) \frac{1}{\sigma}\, d\theta\, d\sigma\, dx \\
&= \int\int \sum_{i=1}^{p} \left\{ \frac{b}{\sigma} - \int \phi p(x|\theta, \sigma)\, d\theta \right\} I(x_i = x_{(1)}) \frac{1}{\sigma}\, d\sigma\, dx \\
&= \int\int\int \frac{1}{\sigma} \left\{ \frac{b}{\sigma} - p(x_{(1)}|\theta_{(1)}, \sigma) \right\} \phi\, d\theta_{(1)}\, d\sigma\, dx_{(1)}. \qquad (4.7)
\end{aligned}$$

Thus, the minimizer of (4.7) is given by

$$\phi_B = \begin{cases} 1, & \text{if } \frac{b}{\sigma} < p(x_{(1)}|\theta_{(1)}, \sigma) \\ 0, & \text{otherwise} \end{cases}$$

or equivalently, substituting the value of b,

$$\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{h^2}{2}\right) < \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_{(1)} - \theta_{(1)})^2}{2\sigma^2}\right).$$

That is

$$\phi_B = \begin{cases} 1, & \text{if } |X_{(1)} - \theta_{(1)}| \leq h\sigma \\ 0, & \text{otherwise.} \end{cases}$$

We find that the rule ϕ_B is a Bayes rule with respect to the right Haar measure. By Lemma 9.2.1 (Robert (2007)) ϕ_B has constant risk. Because its risk is constant the best equivariant rule ϕ_B is minimax.

4.2.1 Unknown σ case

For unknown σ case, put a prior $\pi(\boldsymbol{\theta}, \sigma^2) = \pi(\boldsymbol{\theta}|\sigma^2)\pi(\sigma^2)$

$$\int_{\mathcal{X}} \int_{\sigma} \int_{\Theta|\sigma,\mathcal{X}} L(\boldsymbol{\theta}_s, \phi)\pi(\boldsymbol{\theta}|\mathbf{x}, \sigma^2)\pi(\sigma^2|\mathbf{x})m(\mathbf{x})\, d\boldsymbol{\theta}\, d\sigma^2\, d\mathbf{x}$$
$$= \int_{\mathcal{X}} \int_{\sigma} \int_{\Theta|\mathcal{X},\sigma} \left[b - \pi(\boldsymbol{\theta}_s|\mathbf{x}, \sigma^2)\right] I(\boldsymbol{\theta}_s \in C)\, d\boldsymbol{\theta}_s \pi(\sigma^2|\mathbf{x})\, d\sigma^2 m(\mathbf{x})\, d\mathbf{x},$$

where $m(\mathbf{x})$ is the marginal density of $\mathbf{X}$.

In this case the Bayes rule is

$$\phi_\tau = \begin{cases} 1, & \text{if } \int_0^\infty [\pi(\boldsymbol{\theta}_s|\mathbf{x}, \sigma^2)\, d\sigma^2 - b] \geq 0, \\ 0, & \text{otherwise.} \end{cases}$$

A more common way to deal with unknown σ is to estimate σ^2 with s^2, such that $s^2 \sim \frac{\sigma^2}{\nu}\chi^2_\nu$. In this case, the Bayes rule becomes

$$\phi_\tau = \begin{cases} 1, & \text{if } \pi(\boldsymbol{\theta}_s|\mathbf{x}, s^2) \geq b \int_0^\infty \frac{1}{\sigma^k}\pi(\sigma^2|\mathbf{x}, s^2)\, s\sigma^2, \\ 0, & \text{otherwise.} \end{cases}$$

Similar calculations will show that this set is also minimax.

4.3 Admissibility

We define admissibility for a confidence set in the following way.

Definition 3. *A confidence procedure C^* is admissible, if there doesn't exist a procedure C_1 such that*

(i) $P_\theta(C_1) \geq P_\theta(C^*)$ *for all* $\theta \in \Theta$

(ii) $\nu(C_1) \leq \nu(C^*)$ *for almost all* $x \in R$,

with strict inequality in (i) for some $\theta \in \Theta$, or in (ii) for some subset of R with non-null measure.

A common way to prove inadmissibility of the set is to find a dominating set. This is mostly done by either recentering the set, the resulting set will have the same volume, so in this case you only need to worry about dominating the coverage probability, or for the set with coverage probability $1-\alpha$ trying to shrink the volume.

4.3.1 Case $p > 2$ and $k = 1$

For naive set C_0 coverage probability $P\left(\left(\theta_{(i)} - X_{(i)}\right)^2 \leq h^2\right) < 1 - \alpha$ for $p > 2$ and so in this case

$$R(\theta, C_0) > 2bh - (1 - \alpha), \text{ for } p > 2.$$

Let's consider Qiu & Hwang (2007) intervals as an alternative to C_0. Their intervals are given by

$$C^{QH} = \{\theta_{(1)} : \left|\theta_{(1)} - \hat{M}X_{(1)}\right| \leq v_1(\hat{M})\},$$

where $v_1^2(\hat{M}_1^*) = \hat{M}_1^*(q_1^2 + \log(\hat{M}_1^*))$, $\hat{M}_1^* = \max\left(\hat{M}, M_1\right)$, $\hat{M} = \left(1 - \frac{p-2}{\|\mathbf{X}\|^2}\right)_+$, $M_1 = 1 - \frac{Q_\alpha}{p-2}$ and $q_1 = \Phi^{-1}(1-\alpha/2) = h$, and Q_α is the α-quantile of a χ^2-distribution with p degrees of freedom. In their paper, Qiu & Hwang (2007) prove that, asymptotically, the coverage probability of their intervals $> 1 - \alpha$. Length of their intervals is given by

$$2v_1(\hat{M}_1^*) \leq h.$$

Thus, risk of C^{QH}

$$R(C^{QH}, \theta) \leq 2bh - (1 - \alpha).$$

In Qiu & Hwang (2007) they provide a plot, Figure 2, where they show that QH intervals are always at most as long as the naive intervals. Procedure C^{QH} dominates the naive procedure C_0 asymptotically for all $\boldsymbol{\theta}$. So for $p > 2$, C_0 *is asymptotically inadmissible.*

4.3.2 Almost admissibility

Here we consider the case of selecting one population ($k = 1$) out of two ($p = 2$). In this section we will check sufficient conditions for admissibility from Brown et al.

(1974). Almost admissibility of the set will follow from Assumption II (Brown et al. (1974)) if Assumptions b-d are satisfied. The condition we need to check are the following:

Assumptions:

1. If $\phi_i \in J$ (J is the class of nonrandomized invariant procedures) and $R(\boldsymbol{\theta}, \phi) \to R_0$, then $\phi_i(0) \to \phi_0(0)$. Conversely, if $\phi_i \in J$ and $\phi_i(0) \to \phi_0(0)$ then $\int \left(L(\phi_0(\mathbf{x}), \mathbf{x}) - L(\phi_i(\mathbf{x}), \mathbf{x})\right)^+ m(\mathbf{x})\, d\mathbf{x} \to 0$.
2. $\int \|\mathbf{x}\| L\left(\phi_0(\mathbf{x}), \mathbf{x}\right) m(\mathbf{x})\, d\mathbf{x} < \infty$.
3. $\int_0^\infty d\lambda \left\{ \sup_{\phi \in J} \int_{-\lambda}^{\lambda} \left[L(\phi_0(\mathbf{x}), \mathbf{x}) - L(\phi(\mathbf{x}), \mathbf{x})\right] m(\mathbf{x})\, d\mathbf{x} \right\} < \infty$.

The location invariant procedures (confidence sets) will have the form $\left(X_{(1)} - t, X_{(1)} + t\right)$ for some t. In order to deduce admissibility we verify each of the assumptions above hold.

We will first check Assumption 1. Let $\phi_i \in J$ and $\phi_i(0) \to \phi_0(0)$. Consider the integral

$$\int \left\{L(\phi_0(\mathbf{x}), \mathbf{x}) - L(\phi_i(\mathbf{x}), \mathbf{x})\right\}_+ m(\mathbf{x})\, d\mathbf{x}$$
$$= \int \left\{1 - I(\theta \in C_0) + \nu(C_0) - (1 - I(\theta \in C_i) + \nu(C_i))\right\}_+ m(\mathbf{x})\, d\mathbf{x}.$$

If $\phi_i(0) \to \phi_0(0)$, then $h \to t$, $I(\theta_{(1)} \in C_0) \to I(\theta_{(1)} \in C_i)$, and $\nu(C_0) \to \nu(C_i)$, so that

$$\int \left\{L(\phi_0(\mathbf{x}), \mathbf{x}) - L(\phi_i(\mathbf{x}), \mathbf{x})\right\}^+ m(\mathbf{x})\, d\mathbf{x} \to \int 0 \cdot m(\mathbf{x})\, d\mathbf{x}.$$

The converse holds similarly.

Now consider the second condition. It follows that

$$\int \|\mathbf{x}\| L(\phi_0(\mathbf{x}), \mathbf{x}) m(\mathbf{x})\, d\mathbf{x} = \int \sqrt{x_1^2 + x_2^2} \left[1 - I(\theta_{(1)} \in C_0) + \nu(C_0)\right] m(\mathbf{x})\, d\mathbf{x}$$

$$= \frac{1}{2}\left[(1+2h) \iint \sqrt{x_1^2 + x_2^2} e^{-\frac{x_1^2+x_2^2}{2}}\, dx_1\, dx_2 - \right. \tag{4.8}$$

$$\left. - \iint \sqrt{x_1^2 + x_2^2} I(\theta_{(1)} \in C_0) e^{-\frac{x_1^2+x_2^2}{2}}\, dx_1\, dx_2 \right]. \tag{4.9}$$

For the integral in (4.8) let $x_1 = r\cos\varphi$, $x_2 = r\sin\varphi$, then

$$\begin{aligned}
\iint \sqrt{x_1^2 + x_2^2} e^{-\frac{x_1^2+x_2^2}{2}}\, dx_1\, dx_2 &= \int_0^{2\pi} \int_0^{\infty} r e^{-r^2/2} \cdot r\, dr\, d\varphi \\
&= \int_0^{2\pi} \left[\sqrt{\frac{\pi}{2}}\, \mathrm{erf}\left(\frac{1}{\sqrt{2}}\right) - \frac{1}{\sqrt{e}}\right] d\varphi \\
&= 2\pi \left[\sqrt{\frac{\pi}{2}}\, \mathrm{erf}\left(\frac{1}{\sqrt{2}}\right) - \frac{1}{\sqrt{e}}\right] < \infty,
\end{aligned}$$

where $\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, dt$ is the error function.

For the integral in (4.9)

$$\begin{aligned}
&\iint \sqrt{x_1^2 + x_2^2} I(\theta_{(1)} \in C_0) e^{-\frac{x_1^2+x_2^2}{2}}\, dx_1\, dx_2 \\
&= \iint \sqrt{x_1^2 + x_2^2} \left[\sum_{i=1}^{2} I(x_i - h \le \theta_i \le x_i + h) I(x_i = x_{(1)})\right] e^{-\frac{x_1^2+x_2^2}{2}}\, dx_1\, dx_2 \\
&= \iint \sqrt{x_1^2 + x_2^2} e^{-\frac{x_1^2+x_2^2}{2}} I(x_1 - h \le \theta_1 \le x_1 + h) I(x_1 = x_{(1)})\, dx_1\, dx_2 \\
&\qquad + \iint \sqrt{x_1^2 + x_2^2} e^{-\frac{x_1^2+x_2^2}{2}} I(x_2 - h \le \theta_2 \le x_2 + h) I(x_2 = x_{(1)})\, dx_1\, dx_2 \\
&= \int \int_{\theta_{(1)}-h}^{\theta_{(1)}+h} \sqrt{x_{(1)}^2 + x_{(2)}^2} e^{-\frac{x_{(1)}^2+x_{(2)}^2}{2}}\, dx_{(1)}\, dx_{(2)} \qquad (4.10) \\
&\qquad + \int \int_{\theta_{(2)}-h}^{\theta_{(2)}+h} \sqrt{x_{(1)}^2 + x_{(2)}^2} e^{-\frac{x_{(1)}^2+x_{(2)}^2}{2}}\, dx_{(1)}\, dx_{(2)}. \qquad (4.11)
\end{aligned}$$

Consider, for example, integral in (4.11) and make polar transformation

$$\int_{-\infty}^{\infty} \int_{\theta_{(1)}-h}^{\theta_{(1)}+h} \sqrt{x_{(1)}^2 + x_{(2)}^2} e^{-\frac{x_{(1)}^2+x_{(2)}^2}{2}}\, dx_{(1)}\, dx_{(2)} = \int_0^{2\pi} \int_0^{\infty} r e^{-r^2/2} < \infty.$$

A similar argument holds for the other term in (4.10). Therefore, $\int \|\mathbf{x}\| L(\phi_0(\mathbf{x}), \mathbf{x}) m(\mathbf{x})\, d\mathbf{x} < \infty$.

Finally, the last assumption is equivalent to showing that the expression depends only on the tails of $m(\mathbf{x})$. Brown (1966) shows that if the following is satisfied suppose there exists a $\lambda_0 < \infty$ such that

$$\int_{-\lambda_0}^{\lambda_0} m(\mathbf{x})\, d\mathbf{x} = 1, \tag{4.12}$$

then the third condition holds. This condition is usually easier to check than condition 3. We need to find λ_0 such that (4.12) is satisfied. Consider the following

$$\begin{aligned}\int_{-\lambda_0}^{\lambda_0} m(\mathbf{x})\, d\mathbf{x} &= \int_{-\lambda_0}^{\lambda_0}\int_{-\lambda_0}^{\lambda_0} \frac{1}{2\pi} e^{-\frac{x_1^2+x_2^2}{2}}\, dx_1\, dx_2 \\ &= \left[\Phi(\lambda_0) - \Phi(-\lambda_0)\right]^2.\end{aligned}$$

So now we can just find λ_0 as a solution of

$$\left[\Phi(\lambda_0) - \Phi(-\lambda_0)\right]^2 = 1.$$

Thus, we were able to find λ_0 so that the above expression holds. We have that all of the conditions for admissibility are satisfied, and we can conclude that the usual confidence set ϕ_0 is *almost admissible* for choosing $k = 1$ population out of $p = 2$ populations considered.

4.4 Examples

In this section we consider adapting confidence sets that dominate the naive set in the non-selection problem to selection context. A number of dominating confidence sets have been proposed, for description of some of them see Section 1.2.1. Here we will only look at the recentered at the positive part James-Stein estimator of Hwang & Casella (1982) and empirical Bayes confidence set with variable radius of Casella & Hwang (1983). However, there are no other papers on improving on

the usual confidence set for multivariate mean estimation in the selection context, so there are no available procedures for direct comparison. There are a few works on confidence interval estimation for the selection problem. To be able to compare these confidence intervals with our set directly we can transform our confidence set to conservative confidence interval and see how they perform.

4.4.1 Recentered naive set

Hwang & Casella (1982) proved that recentering the usual naive set at the James-Stein estimator for $p \geq 4$, i.e. the set

$$C_{\delta_+} = \left\{ \boldsymbol{\theta} : \|\boldsymbol{\theta} - \delta_+(\mathbf{X})\|^2 \leq h^2 \right\}, \tag{4.13}$$

where $\delta_+(\mathbf{x}) = \left(1 - \frac{a}{\|\mathbf{X}\|^2}\right)_+ \mathbf{X}$ and $h^2 = \chi^2_{1-\alpha,p}$, leads to domination in coverage probability for all $\boldsymbol{\theta}$ for values of $0 < a \leq a_0$ satisfying

$$\left(\frac{c + (c^2 + a_0)^{1/2}}{\sqrt{a_0}}\right)^{p-3} e^{-c\sqrt{a_0}} = 1. \tag{4.14}$$

The upper bounds on a are less than $p-2$, but Hwang & Casella (1982) provide numerical results to show that coverage probabilities are basically the same as for the case when a satisfies (4.14). Recentered sets are of interest because they have the same volume as the naive sets, and in the non-selection context they provide substantial gains in coverage probability.

Let's consider using similar approach for selection problem, that is we want to use

$$C_{\delta^+} = \left\{ \boldsymbol{\theta}_s : \sum_{i=1}^{k} \left(\theta_{(i)} - \left(1 - \frac{a}{1 + \|\mathbf{X}\|^2}\right)_+ X_{(i)} \right)^2 \leq h^2 \right\} \tag{4.15}$$

with $a = p - 2$ and $h^2 = \chi^2_{1-\alpha,k}$.

Table 4.1: Coverage probabilities for the set (4.15) where $a = p - 2$, $k = 1$.

$\|\boldsymbol{\theta}\|$	$p = 3$	$p = 5$	$p = 7$	$p = 9$	$p = 11$	$p = 13$	$p = 15$	$p = 25$
0	0.923	0.968	0.974	0.985	0.984	0.983	0.988	0.977
2	0.901	0.952	0.965	0.966	0.976	0.981	0.993	0.999
4	0.887	0.919	0.908	0.933	0.955	0.959	0.957	0.982
6	0.905	0.921	0.933	0.931	0.925	0.933	0.941	0.969
8	0.891	0.915	0.916	0.921	0.93	0.923	0.94	0.944
10	0.903	0.909	0.922	0.927	0.923	0.93	0.922	0.92
15	0.913	0.916	0.924	0.908	0.925	0.912	0.9	0.896
20	0.905	0.895	0.92	0.92	0.924	0.92	0.911	0.88
25	0.89	0.893	0.904	0.903	0.921	0.92	0.899	0.879
50	0.899	0.901	0.912	0.899	0.925	0.911	0.909	0.892
100	0.902	0.91	0.901	0.9	0.902	0.9	0.908	0.916
500	0.904	0.901	0.902	0.893	0.905	0.898	0.894	0.896

From Figure 3.1 we can see that the naive set (3.3) does not perform well for all $\boldsymbol{\theta}$ for the case $p > 2$. Via the simulations, we can see that recentered sets (4.15) perform at least as well and in many cases are much better for smaller values of $\|\boldsymbol{\theta}\|$ and higher p.

Table 4.2: Coverage probabilities for the set (4.15) and for the naive set where $a = p - 2$.

		$k = 1$		$k = 2$	
	$\boldsymbol{\theta}$	Recentered	Naive	Recentered	Naive
$p = 3$	$(0, 0, 0)$	0.967	0.936	0.985	0.964
	$(0, 0.25, 0.5)$	0.963	0.920	0.978	0.955
	$(0, 5, 10)$	0.945	0.946	0.950	0.952
	$(0, 0, 2)$	0.946	0.942	0.977	0.947
	$(0, 0, 5)$	0.932	0.952	0.952	0.955
$p = 5$	$(0, 0, 0, 0, 0)$	0.993	0.924	0.985	0.917
	$(0, 0.25, 0.5, 0.75, 1)$	0.983	0.896	0.989	0.93
	$(1, 3, 5, 7, 9)$	0.968	0.964	0.960	0.96
	$(0, 0, 2, 2, 2)$	0.951	0.911	0.977	0.947
	$(0, 0, 5, 5, 5)$	0.962	0.940	0.980	0.968
$p = 10$	$(0, 0, 0, 0, 0, 0, 0, 0, 0, 0)$	0.992	0.770	0.993	0.822
	$(0, 0, 0, 0, 0, 1, 1, 1, 1, 1)$	0.983	0.799	0.995	0.839
	$(0, 0, 0, 0, 0, 5, 5, 5, 5, 5)$	0.947	0.884	0.981	0.93
	$(0.1, 0.2, \ldots, 1)$	0.983	0.768	0.990	0.812
	$(1, 2, \ldots, 10)$	0.955	0.940	0.969	0.946

Table 4.3: Coverage probabilities for the set (4.15) and for the naive set where $a = p - 2$.

		$k = 5$	
	$\boldsymbol{\theta}$	Recentered	Naive
$p = 10$	$(0, 0, 0, 0, 0, 0, 0, 0, 0, 0)$	0.999	0.924
	$(0, 0, 0, 0, 0, 1, 1, 1, 1, 1)$	1	0.946
	$(0, 0, 0, 0, 0, 5, 5, 5, 5, 5)$	0.947	0.951
	$(0.1, 0.2, \ldots, 1)$	0.998	0.947
	$(1, 2, \ldots, 1)$	0.965	0.958
$p = 50$	0	1	0.18
	$\theta_1 = \cdots = \theta_{25} = 0, \theta_{26} = \cdots = \theta_{50} = 1$	0.999	0.372
	$\theta_1 = \cdots = \theta_{25} = 0, \theta_{26} = \cdots = \theta_{50} = 5$	0.964	0.628
	$(0.1, 0.2, \ldots, 5)$	0.976	0.742
	$(1, 2, \ldots, 50)$	0.952	0.952
	$(2, 4, \ldots, 100)$	0.94	0.94
	$(5, 10, \ldots, 250)$	0.957	0.958

Unfortunately, we were not able to prove domination in coverage probability analytically. The way Hwang & Casella (1982) were able to prove that recentered sets have higher coverage probability than the naive sets is by proving the following.

By assuming that for $\|\boldsymbol{\theta}\| < c$, $P_\theta(C_{\delta_+}) \geq P_\theta(C_x^0)$ and since $\lim_{\|\boldsymbol{\theta}\|\to\infty} P_\theta(C_{\delta_+}) = P_\theta(C_x^0)$, a sufficient condition for domination of C_{δ_+} in coverage probability is that

$$\frac{\partial}{\partial\|\boldsymbol{\theta}\|} P_\theta(C_{\delta_+}) \leq 0 \text{ for } \|\boldsymbol{\theta}\| > c. \tag{4.16}$$

They could prove the result by defining a function

$$B_n(\boldsymbol{\theta}) = \int \Phi_n \left(c^2 - \left[\left(1 - \frac{a}{\|\mathbf{x}\|^2} \right) x_{(1)} - \theta_{(1)} \right]^2 \right) p(\mathbf{x}|\boldsymbol{\theta})\, d\mathbf{x}, \tag{4.17}$$

such that $\lim_{n\to\infty} B_n(\boldsymbol{\theta}) = P(C_\theta)$ and $\lim_{n\to\infty} \frac{\partial}{\partial\|\boldsymbol{\theta}\|} B_n(\boldsymbol{\theta}) = \frac{\partial}{\partial\|\boldsymbol{\theta}\|} P(C_\theta)$ for $\|\boldsymbol{\theta}\| > c$.

The derivative of B_n is easier to deal with than that of the coverage probability $P_\theta(C_{\delta_+})$. Derivations in Hwang & Casella (1982) are greatly simplified by assuming $\boldsymbol{\theta} = (\|\boldsymbol{\theta}\|, 0, \ldots, 0)$ since both $P_\theta(C_{\delta_+})$ and B_n depend on $\boldsymbol{\theta}$ only through $\|\boldsymbol{\theta}\|$. For recentered set (4.15) the coverage probability and B_n will depend on both $\|\boldsymbol{\theta}\|$ and

$\boldsymbol{\theta}_s$ which makes calculations more difficult and similar simplification as in Hwang & Casella (1982) is not available.

4.4.2 Empirical Bayes set under normal prior

Casella & Hwang (1983), using an empirical Bayes approach, provide a confidence set, a recentered sphere at the positive part James-Stein estimator, with uniformly smaller volume than the usual confidence set. They give numerical evidence to show that proposed set uniformly dominates the usual set in coverage probability for a wide range of $\boldsymbol{\theta}$ and p, but analytical dominance results were not provided. The derivation is based on empirical Bayes approach under the loss (4.1). Their resulting set updated to take into account selection is

$$C_\delta^E = \{\theta_{(1)} : |\theta_{(1)} - \delta_+|^2 \leq v_E^2(\|\mathbf{X}\|)\}, \tag{4.18}$$

where $\delta_+ = \left(1 - \frac{p-2}{\|\mathbf{X}\|^2}\right)_+ X_{(1)}$, has variable radius

$$v_E^2(\|\mathbf{X}\|) = \begin{cases} \left(1 - \frac{p-2}{h^2}\right)\left[h^2 - p\log\left(1 - \frac{p-2}{h^2}\right)\right] & \text{if } \|\mathbf{X}\| \leq h, \\ \left(1 - \frac{p-2}{|\mathbf{X}|^2}\right)\left[h^2 - p\log\left(1 - \frac{p-2}{\|\mathbf{X}\|^2}\right)\right] & \text{if } \|\mathbf{X}\| > h. \end{cases}$$

Hwang & Casella (1982) proved that recentered usual set at the positive James Stein estimator ($p \geq 4$) uniformly dominates the usual confidence set, and therefore is minimax. For the performance of the set (4.18) see Table 4.4.

4.4.3 Empirical Bayes set under the horseshoe prior

Here we will follow the construction of confidence sets similarly to Casella & Hwang (1983) approach for selection problem using horseshoe prior (2.37). In this section we assume as before that τ is estimated empirically by (2.38).

First, we need to find the Bayes rule for selection under horseshoe prior. The Bayes rule under the loss (4.1) for a general prior $\pi(\theta)$ is given by (4.2). From van der Pas et al. (2014) we know that under the horseshoe prior, posterior mean $\theta_{(i)}|x_{(i)}, \tau$ is normally distributed with mean $\delta_g(x_{(i)})$ given by (2.40) and variance

$$\begin{aligned} Var(\theta_{(i)}|\mathbf{x}) &= \frac{1}{x_{(i)}}\delta_g - \left(\delta_g - x_{(i)}\right)^2 + x_{(i)}^2 \frac{\int_0^1 t^{-\frac{1}{2}}(1-t)^2 \frac{1}{\tau^2+(1-\tau^2)t} e^{\frac{x_{(i)}^2}{2}t}\, dt}{t^{-\frac{1}{2}} \frac{1}{\tau^2+(1-\tau^2)t} e^{\frac{x_{(i)}^2}{2}t}\, dt} \\ &= \sum_{j=1}^{p} \left[\frac{1}{x_j}\delta_g - (\delta_g - x_j)^2 + x_j^2 \frac{8\Phi_1\left(\frac{1}{2}, 1, \frac{7}{2}; \frac{x_j^2}{2}, 1-\frac{1}{\tau^2}\right)}{15\Phi_1\left(\frac{1}{2}, 1, \frac{3}{2}; \frac{x_j^2}{2}, 1-\frac{1}{\tau^2}\right)} \right] I(x_j = x_{(i)}), \end{aligned} \tag{4.19}$$

where

$$\Phi_1(\alpha, \beta, \gamma, x, y) = \sum_{m,n=0}^{\infty} \frac{(a)_{m+n}(b)_m}{(c)_{m+n} m! n!} x^m y^n \text{ for } |x| < 1 \tag{4.20}$$

is the degenerate hypergeometric function of two variables or Humbert series.

Using the facts above we can write the generalized Bayes set for estimating k selected parameters in the following way

$$\begin{aligned} C_X^B &= \left\{ \boldsymbol{\theta} : \prod_{i=1}^{k} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(\theta_{(i)} - \delta_g)^2}{2\mathrm{Var}(\theta_{(i)}|\mathbf{x})} \right) \geq b \right\} \\ &= \left\{ \boldsymbol{\theta} : \sum_{i=1}^{k} \frac{(\theta_{(i)} - \delta_g)^2}{\mathrm{Var}(\theta_{(i)}|\mathbf{x})} \leq h^2 \right\}, \end{aligned} \tag{4.21}$$

where for non-adjusted for selection sets $h^2 = \chi^2_{1-\alpha,k}$.

Table 4.4 provides coverage probabilities for the confidence sets (4.18) and (4.21). Empirical Bayes set under the normal prior (4.18) mostly performs well except for the sparse scenarios with small signal. Empirical Bayes set under the horseshoe prior performs well for some setups but for other gives very low coverage probabilities.

These sets need to be adjusted to selection further. We expect that similar technique as in section on approximating the coverage probability can be used here

to derive an approximation for the coverage probability of the generalized Bayes confidence sets, which can be used to adjust (4.21) to guarantee coverage probability to be at least $1 - \alpha$.

Table 4.4: Coverage probabilities for the sets (4.18) (EB normal) and (4.21) (EB horseshoe), $p = 5$.

	$k = 3$	
$\boldsymbol{\theta}$	EB normal	EB horseshoe
$(0, 0, 0, 0, 0)$	0.97	1
$(5, 5, 5, 5, 5)$	0.98	0.99
$(0, 0.25, 0.5, 0.75, 1)$	0.98	0.66
$(0, 0, 0, 2, 2)$	0.95	0.54
$(0, 0, 0, 5, 5)$	0.93	0.92
$(1, 2, 3, 4, 5)$	0.98	0.93

Table 4.5: Coverage probabilities for the sets (4.18) (EB normal) and (4.21) (EB horseshoe), $p = 10$.

	$k = 3$		$k = 5$	
$\boldsymbol{\theta}$	EB normal	EB horseshoe	EB normal	EB horseshoe
$(0, \ldots, 0)$	-	1	0.98	1
$(5, \ldots, 5)$	0.88	0.97	0.93	0.99
$(0, \ldots, 0, 1, \ldots, 1)$	-	0.45	0.97	0.23
$(0, \ldots, 0, 5, \ldots, 5)$	0.98	0.99	0.97	0.94
$(0.1, 0.2, \ldots, 1)$	-	0.79	0.96	0.47
$(1, 2, \ldots, 1)$	0.95	0.94	0.94	0.98

Table 4.6: Coverage probabilities for the sets (4.18) (EB normal) and (4.21) (EB horseshoe), $p = 50$.

	$k = 5$		$k = 10$	
$\boldsymbol{\theta}$	EB normal	EB horseshoe	EB normal	EB horseshoe
$(0, \ldots, 0)$	-	1	-	1
$(5, \ldots, 5)$	0.68	0.68	0.77	0.86
$(0, \ldots, 0, 1 \ldots, 1)$	1	0.29	-	0.06
$(0, \ldots, 0, 5 \ldots, 5)$	0.96	0.91	1	1
$(0.1, 0.2, \ldots, 5)$	0.99	0.97	1	0.99
$(1, 2, \ldots, 50)$	0.95	-	0.96	-

CHAPTER 5

DISCUSSION

5.1 Conclusions

As high dimensional data becomes more common these days, selection always comes into play when we only want to estimate a subset of features or parameters of the model. It was known since the work of Stein (1956) that the naive estimate of the selected mean is biased, and some improved point estimates have been proposed. But only minimaxity of the naive estimate was considered. Here we explored the admissibility of the naive estimates of the selected means. In Chapter 2 we were able to prove that $X_{(1)}$ is admissible for estimating $\theta_{(1)}$ for $p \leq 3$. For the case $p \geq 3$ no analytical inadmissibility results provided, but we conjecture that it is inadmissible. And as a consequence, we proved that generalized Bayes estimate of $\theta_{(1)}$ given in (2.21) is admissible for $p \geq 4$. We also provide some results concerning the admissibility of the generalized Bayes estimator of k selected parameters. Resulting estimator (2.40) under the horseshoe prior performs well under the sparsity scenario compared to some other estimators.

With a redeveloped interest in the selection problem in the last ten years, a lot was done for point estimation, some of the works proposed how to deal with confidence intervals, but there are no papers that we know of that dealt with confidence sets after selection. Analytically, confidence set estimation after selection is much harder to deal with, since all of the calculations will involve integration over a k-dimensional sphere instead of a repeated integrals for confidence intervals. We were not able to provide a theoretical domination results of the considered sets over the naive confidence set under the loss (4.1). So even though the naive set does not maintain nominal coverage probability of $1 - \alpha$, because it uses unadjusted sets

which are independent of the total number of populations p, it is very small, and thus it is hard to dominate under the loss (4.1). We did provide the confidence sets with guaranteed coverage probability of $1-\alpha$ that are in general wider than the naive sets. The confidence sets that do take selection into account are expected to be wider to accommodate for the variability of the selection mechanism.

We also proved minimaxity of the naive set for any p, total number of populations, and k, number of selected populations, for a less restricted definition of minimaxity that does not have a coverage probability constraint. For a stronger definition of minimaxity which requires a confidence set to be $1-\alpha$, the naive set is only admissible for the case $p = 2$ and $k = 1$.

A lot of the proofs for selection can be obtained by adapting the proof for no selection point estimation and confidence sets problems. After we condition on the data, we only need to look at the terms corresponding to the selected population.

The following table provides a summary of the results obtained in the literature and in this dissertation.

Table 5.1: Summary of minimaxity and admissibility results from selection (S) and non-selection (NS).

	Minimaxity	**Admissibility**
NS Point Estimate	any p, Stein (1964)	$p < 3$, Stein (1956)
S Point Estimate	$p < 3$	$X_{(1)}$ is admissible for $p \leq 3$
	Sackrowitz & Samuel-Cahn (1986)	δ_g (2.27) is admissible for $p \geq 4$
NS Confidence Sets	Casella & Hwang (1991)	Joshi (1969) for $p \leq 2$
S Confidence Sets	$p = 2$ under Definition 2	almost admissible for $p = 2$
	any p under Definition 1	

5.2 Future work

5.2.1 Admissibility of the naive confidence set for general p and $k = 1$

For the case of general p and $k = 1$ we suspect that the naive confidence set is inadmissible. We would want to use some analog of Blyth method to the confidence sets under the loss (4.1).

Need to proof result like Theorem A.1 in Maruyama (2009). We conjecture it should be along the lines of: C_g is admissible under the priors $G(\|\boldsymbol{\theta}\|)$ if

$$\nu(C_g) - \nu(C_{g_n}) \to 0 \text{ and,}$$
$$I(\theta_{(1)} \in C_g) - I(\theta_{(1)} \in C_{g_n}) \to 0,$$

under some conditions (which are probably similar to conditions from Theorem A.1)

Consider the following sequence of priors (2.17) following Maruyama (2009). In the paper Maruyama (2009) proves that admissibility can be proven under harmonic priors of the form $G(\|\boldsymbol{\theta}\|) = \|\boldsymbol{\theta}\|^{p-2}$. Here we consider independent priors on θ_i's and we are only interested in the case $p = 2$, so in here we consider $G(|\theta_i|) = 1$. Under the priors $g_i = G(|\theta_i|)$, the Bayes rule is given by

$$C_g = \left\{ \theta_{(1)} : \frac{\phi(x_{(1)} - \theta_{(1)}) G(|\theta_{(1)}|)}{\int_\Theta \phi(x_{(1)} - \theta_{(1)}) G(|\theta_i|)\, d\theta_{(1)}} \right\}$$
$$\Rightarrow C_g = \left\{ \theta_{(1)} : \frac{\phi(x_{(1)} - \theta_{(1)})}{\int_\Theta \phi(x_{(1)} - \theta_{(1)})\, d\theta_{(1)}} \right\}$$
$$\Rightarrow C_g = \left\{ \theta_{(1)} : \phi(x_{(1)} - \theta_{(1)}) \geq h \right\} = C_0.$$

The Bayes rule under the sequence of priors g_n is given by

$$C_{g_n} = \left\{ \theta_{(1)} : \frac{\phi(x_{(1)} - \theta_{(1)}) G(|\theta_{(1)}|) H_n(|\theta_{(1)}|)}{\int_\Theta \phi(x_{(1)} - \theta_{(1)}) G(|\theta_i|) H_n(|\theta_{(1)}|)\, d\theta_{(1)}} \right\}$$
$$\Rightarrow C_{g_n} = \left\{ \theta_{(1)} : \frac{\phi(x_{(1)} - \theta_{(1)}) H_n(|\theta_{(1)}|)}{\int_\Theta \phi(x_{(1)} - \theta_{(1)}) H_n(|\theta_{(1)}|)\, d\theta_{(1)}} \right\}.$$

From above, for admissibility we need the following condition to be satisfied as $n \to \infty$

$$\frac{\pi(\theta_{(1)})}{\int \phi(x_{(1)} - \theta_{(1)}) \pi(\theta_{(1)})\, d\theta_{(1)}} \to 1.$$

Under the priors (2.17)

$$\begin{aligned}
\lim_{n\to\infty} \frac{\pi(\theta_{(1)})}{\int \phi(x_{(1)} - \theta_{(1)}) \pi(\theta_{(1)})\, d\theta_{(1)}} &= \lim_{n\to\infty} \frac{H_n(|\theta_{(1)}|)}{\int \phi(x_{(1)} - \theta_{(1)}) H_n(|\theta_{(1)}|)\, d\theta_{(1)}} \\
&= \frac{\lim_{n\to\infty} H_n(|\theta_{(1)}|)}{\lim_{n\to\infty} \int \phi(x_{(1)} - \theta_{(1)}) H_n(|\theta_{(1)}|)\, d\theta_{(1)}} \\
&= \frac{1}{\int \phi(x_{(1)} - \theta_{(1)}) \lim_{n\to\infty} H_n(|\theta_{(1)}|)\, d\theta_{(1)}} \\
&= 1,
\end{aligned}$$

since $\lim_{n\to\infty} H_n(t) = 1$ (Theorem 2.2 in Maruyama (2009)).

As long as we have conditions to guarantee that above is sufficient for admissibility, we can prove admissibility of the usual set in this case. Unfortunately, it is not clear how to generalize Blyth method for confidence sets under the loss (4.1).

5.2.2 The confidence report problem

As was mentioned in the introduction, when we are estimating a confidence set, we want estimated set to have good conditional properties. Casella (1992) gave an overview of how conditional (post-data) inference was developed. A frequentist

solution to the problem of confidence set estimation usually yields a set that has a minimum coverage guarantee. Unfortunately, with frequentist solutions we do not have a guarantee that post-experimental coverage will be at the nominal level. Extensive research was done for constructing confidence intervals with good conditional performance. Robinson (1979a) was one of the first people to discuss this problem, and he considered conditional inference in terms of betting strategies and proved that absence of semi relevant betting procedures implies admissibility with respect to squared error loss. Using this approach Robinson (1979b) showed that the usual interval estimates for the case $p = 1$ are admissible and are not admissible for $p \geq 5$.

The alternative to frequentist solutions, Bayes procedures are conditionally acceptable (Robinson (1979a), Pierce (1973)), so you only need to prove that the estimated confidence set is a a Bayes credible set with coverage probability of at least $1 - \alpha$.

An alternative to deriving conditionally accepted procedures is to consider "estimated confidence" approach introduced in Kiefer (1977). Following Neyman-Pearson theory, it is common for a confidence set $C(\mathbf{x})$ to report the minimum coverage probability $\gamma = 1 - \alpha$. This constant coverage estimator is often inadmissible estimator of the coverage function $I(\theta \in C_x)$. Also for example we know that sets recentered at the positive part James-Stein estimator dominate the naive set (Hwang & Casella (1982)). Both of these sets have the same minimum coverage probability $1-\alpha$, but if we report $1-\alpha$ for the recentered set, we do not provide any information about how better this set actually is. Therefore, it is better to report instead of the constant value $1 - \alpha$, data dependent (or conditional on the data) estimate $1 - \alpha(x)$ of the coverage probability. Brown & Hwang (1990) considered estimating $I(\boldsymbol{\theta} \in C_x)$ directly under the squared error loss

$$L(\boldsymbol{\theta}, \gamma) = (\gamma(\mathbf{x}) - I(\boldsymbol{\theta} \in C_x))^2 . \tag{5.1}$$

They proved that the constant coverage estimator is admssible for $p \leq 4$. They also

proved admissibility of the generalized Bayes estimator under loss (5.1) under some conditions.

Hwang & Brown (1991) considered supplementing estimated confidence approach with frequentist validity constraint. They defined it in the following way. Confidence $\gamma = 1 - \alpha$ is a valid confidence for C_x if

$$E_{\boldsymbol{\theta}} I(\boldsymbol{\theta} \in C_x) \geq \gamma, \text{ for any } \boldsymbol{\theta}. \tag{5.2}$$

Under this validity constraint they proved that the usual constant coverage probability estimator is admissible for all p. They call it validity admissibility.

Following Brown & Hwang (1990), Wang (1998) was also interested in admissibility of the usual estimate of the coverage function, but he considered confidence intervals for a randomly chosen linear combination of θ_i's. He proved that $1 - \alpha$ is an admissible estimator of $I(\mathbf{w}'\boldsymbol{\theta} \in C_{\mathbf{x},\mathbf{w}})$ for $p \leq 4$ under squared error loss. Here $\mathbf{X}$ and $\mathbf{W}$ are independent. In the companion paper Wang (1999) proved that if $\mathbf{W}$ is random, it is inadmissible for $p \geq 5$, and is admissible if $\mathbf{W}$ is fixed. For the case $p \geq 5$, Wang (2005) provided improved confidence report that dominates the constant coverage estimator.

Remark: For the selection problem, we could consider similar setup with $\mathbf{W} := \{w_i\}_{i=1}^p = \sum_{i=1}^p I(X_i = X_{(j)})$. Here $\mathbf{W}$ and $\mathbf{X}$ are not independent.

George & Casella (1994) considered an empirical Bayes approach to estimating the confidence report and provided estimates of the coverage that dominate constant estimate $1 - \alpha$.

It would be interesting to consider estimating confidence report of the confidence sets proposed in Section 3.2.2.

BIBLIOGRAPHY

Bechhofer, R. (1954). A single-sample multiple decision procedure for ranking means of normal populations with known variances. *Annals of Mathematical Statistics*, *25*(1), 16–39.

Benjamini, Y., & Yekutieli, D. (2005). False discovery rate adjusted multiple confidence intervals for selected parameters. *Journal of the American Statistical Association*, *100*(469), 71–81.

Berger, J. O. (1976a). Admissibility results for generalized bayes estimators of coordinates of a location vector. *The Annals of Statistics*, *4*(2), 334–356.

Berger, J. O. (1976b). Inadmissibility results for generalized Bayes estimators of coordinates of a location vector. *Annals of Statistics*, *4*(2), 302–333.

Berger, J. O. (1976c). Tail minimaxity in location vector problems and its applications. *The Annals of Statistics*, *4*(1), 33–50.

Berger, J. O., & Strawderman, W. E. (1996). Choice of hierarchical priors: admissibility in estimation of normal means. *The Annals of Statistics*, *24*(3), 931–951.

Blumenthal, S. (1970). Interval estimation of the normal mean subject to restrictions, when the variance is known. *Naval Research Logistics Quarterly*, *17*(4), 485–505.

Brown, L. (1966). On the admissibility of invariant estimators of one or more location parameters. *The Annals of Mathematical Statistics*, *37*(5), 1087–1136.

Brown, L. (1979). A heuristic method for determining admissibility of estimators with applications. *Annals of Statistics*, *7*(5), 960–994.

Brown, L., Fox, M., et al. (1974). Admissibility in statistical problems involving a location or scale parameter. *The Annals of Statistics*, *2*(4), 807–814.

Brown, L., & Hwang, J. (1990). Admissibility of confidence estimators. In M.-T. Chao, & P. E. Cheng (Eds.) *Proceedings of the 1990 Taipei Symposium in Statistics*, (pp. 1–10). Institute of Statistical Science, Academia Sinica.

Brown, L., & Hwang, J. T. (1982). A unified admissibility proof. In *Statistical Decision Theory and Related Topics III*, vol. 1, (pp. 205–230). Academic Press New York.

Carvalho, C. M., Polson, N. G., & Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, *97*(2), 465–480.

Casella, G. (1992). Conditional inference from confidence sets. In *Lecture Notes-Monograph Series: Essays in Honor of D.Basu*, vol. 17, (pp. 1–12). Institute of Mathematical Statistics.

Casella, G., & Hwang, J. (1991). Evaluating confidence sets using loss functions. *Statistica Sinica*, *1*, 159–173.

Casella, G., & Hwang, J. T. (1983). Empirical bayes confidence sets for the mean of a multivariate normal distribution. *Journal of the American Statistical Association*, *78*(383), 688–698.

Casella, G., Hwang, J. T., & Robert, C. (1993). A paradox in decision-theoretic interval estimation. *Statistica Sinica*, *3*(1), 141155.

Cohen, A., & Sackrowitz, H. (1982). Estimating the mean of the selected population. In *Third Purdue Symposium on Statistical Decision Theory and Related Topics*. New York: Academic Press.

Cohen, A., & Strawderman, W. E. (1973). Admissible confidence interval and point estimation for translation or scale parameters. *The Annals of Statistics*, *1*(3), 545–550.

Efron, B. (2006). Minimum volume confidence regions for a multivariate normal mean vector. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *68*(4), 655–670.

Efron, B. (2011). Tweedies formula and selection bias. *Journal of the American Statistical Association*, *106*(496), 1602–1614.

Farrell, R. H. (1964). Estimators of a location parameter in the absolutely continuous case. *The Annals of Mathematical Statistics*, *35*(3), 949–998.

Fuentes, C., Casella, G., & Wells, M. T. (2014). Interval estimation for the mean of the selected populations. *Submitted*.

George, E. I., & Casella, G. (1994). An empirical bayes confidence report. *Statistica Sinica*, *4*, 617–638.

Gibbons, J. D., Olkin, I., & Sobel, M. (1999). *Selecting and Ordering Populations: a New Statistical Methodology*. SIAM.

Gradshteyn, I., & Ryzhik, I. (2007). *Table of Integrals, Series, and Products*. Elsevier, 7th ed.

Gupta, S. S. (1956). *On a decision rule for a problem in ranking means*. Ph.D. thesis, University of North Carolina at Chapel Hill.

Gupta, S. S., & Miescke, K. J. (1986). On the problem of finding the largest normal mean under heteroscedasticity. Tech. rep., DTIC Document.

Gupta, S. S., & Panchapakesan, S. (1979). *Multiple decision procedures: theory and methodology of selecting and ranking populations*, vol. 44. Siam.

Gupta, S. S., & Panchapakesan, S. (1985). Subset selection procedures: review and assessment. *American Journal of Mathematical and Management Sciences*, *5*(3-4), 235–311.

Guttman, I., & Tiao, G. (1964). A Bayesian approach to some best population problems. *Annals of Mathematical Statistics*, *35*(2), 825–835.

He, K. (1992). Parametric empirical bayes confidence intervals based on james-stein estimator. *Statistics & Risk Modeling*, *10*(1-2), 121–132.

Hwang, J. (1993). Empirical Bayes estimation for the means of the selected populations. *Sankhyā: The Indian Journal of Statistics, Series A*, *55*(2), 285–304.

Hwang, J., & Brown, L. (1991). Estimated confidence under the validity constraint. *The Annals of Statistics*, *19*(4), 1964–1977.

Hwang, J. T., & Casella, G. (1982). Minimax confidence sets for the mean of a multivariate normal distribution. *The Annals of Statistics*, *10*(3), 868–881.

Joshi, V. (1967). Inadmissibility of the usual confidence sets for the mean of a multivariate normal population. *The Annals of Mathematical Statistics*, *38*(6), 1868–1875.

Joshi, V. (1969). Admissibility of the usual confidence sets for the mean of a univariate or bivariate normal population. *The Annals of Mathematical Statistics*, *40*(3), 1042–1067.

Kabaila, P. (2011). Admissibility of the usual confidence interval for the normal mean. *Statistics & Probability Letters*, *81*(3), 352–359.

Kiefer, J. (1977). Conditional confidence statements and confidence estimators. *Journal of the American Statistical Association*, *72*(360a), 789–808.

Krylov, V. I., & Stroud, A. H. (2006). *Approximate Calculation of Integrals*. Courier Corporation.

Lee, J. D., Sun, D. L., Sun, Y., & Taylor, J. E. (2014). Exact post-selection inference, with application to the lasso. *arXiv preprint arXiv:1311.6238*.

Lehmann, E. L., & Casella, G. (2003). *Theory of Point Estimation*. Springer Science & Business Media.

Liese, F., & Miescke, K.-J. (2008). *Statistical decision theory: estimation, testing, and selection*. Springer Science & Business Media.

Martin, R., Walker, S. G., et al. (2014). Asymptotically minimax empirical bayes estimation of a sparse normal mean vector. *Electronic Journal of Statistics*, *8*(2), 2188–2206.

Maruyama, Y. (2009). An admissibility proof using an adaptive sequence of smoother proper priors approaching the target improper prior. *Journal of Multivariate Analysis*, *100*(8), 1845–1853.

Meeden, G., & Vardeman, S. (1985). Bayes and admissible set estimation. *Journal of the American Statistical Association*, *80*(390), 465–471.

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

Mustard, D. (1964). Numerical integration over the -dimensional spherical shell. *Mathematics of Computation*, *18*(88), 578–589.

Owen, D. B. (1980). A table of normal integrals. *Communications in Statistics-Simulation and Computation*, *9*(4), 389–419.

Pierce, D. A. (1973). On some difficulties in a frequency theory of inference. *The Annals of Statistics*, *1*(2), 241–250.

Qiu, J., & Hwang, J. (2007). Sharp simultaneous intervals for the means of selected populations with application to microarray data analysis. *Biometrics*, *63*(3), 767–776.

Reid, S., & Tibshirani, R. (2014). Post selection point and interval estimation of signal sizes in gaussian samples. *arXiv preprint arXiv:1405.3340*.

Robert, C. (2007). *The Bayesian Choice: from Decision-Theoretic Foundations to Computational Implementation*. Springer Science & Business Media.

Robinson, G. (1979a). Conditional properties of statistical procedures. *The Annals of Statistics*, *7*(4), 742–755.

Robinson, G. (1979b). Conditional properties of statistical procedures for location and scale parameters. *The Annals of Statistics*, *7*(4), 756–771.

Sackrowitz, H., & Samuel-Cahn, E. (1986). Evaluating the chosen population: a Bayes and minimax approach. In *Lecture Notes-Monograph Series. Adaptive Statistical Procedures and Related Topics*, vol. 8, (pp. 386–399).

Samworth, R. (2005). Small confidence sets for the mean of a spherically symmetric distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(3), 343–361.

Saxena, K., & Tong, Y. (1969). Interval estimation of the largest mean of k normal populations with known variances. *Journal of the American Statistical Association*, *64*(325), 296–299.

Simon, N., & Simon, R. (2013). On estimating many means, selection bias, and the bootstrap. *arXiv preprint arXiv:1311.3709*.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, (pp. 197–206).

Stein, C. (1964). Contribution to the discussion of Bayesian and non-Bayesian decision theory. *Handout from the Institute of Mathematical Statistics Meeting*.

Tseng, Y.-L., Brown, L., et al. (1997). Good exact confidence sets for a multivariate normal mean. *The Annals of Statistics*, *25*(5), 2228–2258.

van der Pas, S., Kleijn, B., van der Vaart, A., et al. (2014). The horseshoe estimator: posterior concentration around nearly black vectors. *Electronic Journal of Statistics*, *8*(2), 2585–2618.

Venter, J. (1988). Confidence bounds based on the largest treatment mean. *South African Journal of Science*, *84*, 340–342.

Venter, J., & Steel, S. (1991). Estimation of the mean of the population selected from k populations. *Journal of Statistical Computation and Simulation*, *38*(1-4), 1–14.

Wang, H. (1998). Admissibility of the constant-coverage probability estimator for estimating the coverage function of certain confidence interval. *Statistics & Probability Letters*, *36*(4), 365–372.

Wang, H. (1999). Brown's paradox in the estimated confidence approach. *Annals of Statistics*, *27*(2), 610–626.

Wang, H. (2005). Improved confidence estimators for confidence sets of location parameters. *Journal of Statistical Planning and Inference*, *128*(1), 95–107.

Winkler, R. L. (1972). A decision-theoretic approach to interval estimation. *Journal of the American Statistical Association*, *67*(337), 187–191.

Xu, J. (2007). A closed form for the harmonic-prior bayes estimator with associated confidence sets for the mean of a multivariate normal distribution. *Dissertation, University of Pennsylvania*.

Zhao, Z., & Hwang, G. J. (2012). Empirical Bayes false coverage rate controlling confidence intervals. *Journal of the Royal Statistical Society: Series B*, *74*(5), 871–891.