Optimal estimation of the null distribution in large-scale inference

Subhodh Kotekal[1] and Chao Gao[2]

University of Chicago

Abstract

The advent of large-scale inference has spurred reexamination of conventional statistical thinking. In a series of highly original articles, Efron persuasively illustrated the danger for downstream inference in assuming the veracity of the theoretical null distribution. However intimidating in other contexts, the large-scale setting is to the statistician's benefit here. There is now potential to estimate, rather than assume, the null distribution. In a Gaussian model for n many z-scores with at most $k < \frac{n}{2}$ nonnulls, Efron suggests estimating the location and scale parameters of the null distribution. Placing no assumptions on the nonnull effects, the statistical task can be viewed as a robust estimation problem. However, the best known robust estimators fail to be consistent in the regime $k \asymp n$ which is especially relevant in large-scale inference. The failure of estimators which are minimax rate-optimal with respect to other formulations of robustness (e.g. Huber's contamination model) might suggest the impossibility of consistent estimation in this regime and, consequently, a major weakness of Efron's suggestion. A sound evaluation of Efron's model thus requires a complete understanding of consistency.

We sharply characterize the regime of k for which consistent estimation is possible and further establish the minimax estimation rates. It is shown consistent estimation of the location parameter is possible if and only if $\frac{n}{2} - k = \omega(\sqrt{n})$, and consistent estimation of the scale parameter is possible in the entire regime $k < \frac{n}{2}$. Faster rates than those in Huber's contamination model are achievable by exploiting the Gaussian character of the data. The minimax upper bound is obtained by considering estimators based on the empirical characteristic function. The minimax lower bound involves constructing two marginal distributions whose characteristic functions match on a wide interval containing zero. The construction notably differs from those in the literature by sharply capturing a scaling of $n - 2k$ in the minimax estimation rate of the location.

1 Introduction

Consider the model

$$X_j = \theta + \gamma_j + \sigma Z_j \tag{1}$$

where $\theta \in \mathbb{R}$ is the location parameter of interest, $\gamma_1, ..., \gamma_n \in \mathbb{R}$ are the unknown effects, $Z_1, ..., Z_n \overset{iid}{\sim} N(0,1)$ are the noise variables, and $\sigma > 0$ is the scale parameter. We are primarily concerned with estimation of θ and σ^2 given observations from (1). We denote the joint distribution of $\{X_j\}_{j=1}^n$ by $P_{\theta,\gamma,\sigma}$ when the data are given by (1); the expectation with respect to $P_{\theta,\gamma,\sigma}$ is denoted by $E_{\theta,\gamma,\sigma}$. Of course, the parameters of interest are not identifiable in (1) as written. To ensure identifiability, we assume a limited number of γ_j are nonzero, that is, $\sum_{j=1}^n \mathbb{1}_{\{\gamma_j \neq 0\}} \leq k < \frac{n}{2}$.

The model (1) is a natural choice in the very typical setting in which a fraction of the data points exhibit perturbed values ("signal" or "corruption" depending on the context) against an unknown background level. Indeed, in investigating this typical statistical setting by considering models similar

---

to (1), the literatures of robust statistics and large-scale inference have witnessed vigorous research activity.

1.1 Robust statistics

An immediate connection between model (1) and robust statistics can be seen by simply conceptualizing X_j as an inlier if $\gamma_j = 0$ and as an outlier otherwise. In the parlance of robust statistics, nonzero γ_j might be referred to as "corruption" or "contamination".

Of course, the robust statistics literature has not studied model (1) exclusively. To describe one foundational backdrop (among others), consider a dataset $X_1, ..., X_n$ in which

$$X_j \sim \begin{cases} N(\theta, \sigma^2) & \text{if } j \in \mathcal{I}, \\ contamination & \text{if } j \in \mathcal{O}. \end{cases}$$

Here, $\mathcal{I}$ denotes the inliers from which the location parameter is to be estimated, and $\mathcal{O}$ with $|\mathcal{O}| \leq k < \frac{n}{2}$ denotes the outliers.

With this backdrop in place, researchers stipulate various models for the contamination process and investigate its effects on statistical tasks. A popular choice is an adversarial contamination process in which an adversary has the opportunity to inspect a dataset of n independent and identically distributed samples from the inlier distribution $N(\theta, \sigma^2)$, make any desired modifications, and pass on the modified dataset to the statistician. Another popular choice with a long history in statistics is Huber's contamination model, proposed by Huber in a highly influential article [32]. A variant is given by

$$X_j \overset{ind}{\sim} \begin{cases} N(\theta, \sigma^2) & \text{if } j \in \mathcal{I}, \\ \delta_{\eta_j} & \text{if } j \in \mathcal{O}. \end{cases} \tag{2}$$

Here, the sets $\mathcal{I}$ and $\mathcal{O}$ are unknown but fixed in advance to the generation of the data. The contamination points η_j are unknown and arbitrary, but also fixed in advance to the data generation. From a minimax perspective, this variant is closely related to the usual Bayes formulation of Huber's contamination model [32], that is,

$$X_1, ..., X_n \overset{iid}{\sim} (1 - \varepsilon)N(\theta, \sigma^2) + \varepsilon Q, \tag{3}$$

where $\varepsilon = \frac{k}{n}$ is the contamination level and Q is an unknown, arbitrary probability distribution on $\mathbb{R}$. That Q can be any probability distribution on $\mathbb{R}$ is due to the fact that the η_j can be any values in $\mathbb{R}$. In a minimax context, the conclusions about in-probability estimation rates in the frequentist version (2) and the Bayes version (3) are the same [13, 20].

Basic statistical intuition correctly asserts sample average is not optimal due to its sensitivity to outliers. It turns out the instinct to reach for sample median is the right one as established in [13], where the minimax rate of estimation was derived and applies to both Bayes (3) and frequentist (2) contexts. In particular, they showed that if $\varepsilon < \frac{1}{5}$, then

$$| \operatorname{median}(X_1, ..., X_n) - \theta|^2 \lesssim \sigma^2 \left(\frac{1}{n} + \varepsilon^2 \right) \asymp \sigma^2 \left(\frac{1}{n} + \frac{k^2}{n^2} \right)$$

with high probability. A matching lower bound (up to universal constants) was obtained, thus establishing the rate-optimality of sample median.

Returning to (1), the situation seems slightly, yet apparently different from that in Huber's contamination model. Rather than outlier observations being set to arbitrary unknown values, outliers from (1) exhibit *mean shifts*. Mean shift contamination has been extensively studied in the robust statistics literature, mostly in a regression context. To elaborate, consider the regression model for data $\{(x_j, Y_j)\}_{j=1}^n$,

$$Y_j = \langle x_j, \beta \rangle + \gamma_j + \sigma Z_j, \tag{4}$$

where $\beta, x_1, ..., x_n \in \mathbb{R}^p$ and $Z_1, ..., Z_n \overset{iid}{\sim} N(0, 1)$. The responses are contaminated by the nonzero mean shifts $\gamma_j \in \mathbb{R}$, and there are at most k nonzero of them. Evidently, the location model (1) is a special case of (4). To the best of our knowledge, the first appearances of this mean shift regression model were in [53, 30, 45]. Gannaz [30], noting $\gamma = (\gamma_1, ..., \gamma_n) \in \mathbb{R}^n$ is a k-sparse vector, suggested what is now a basic impulse for many statisticians,

$$(\hat{\beta}, \hat{\gamma}) = \underset{\substack{\beta \in \mathbb{R}^p \\ \gamma \in \mathbb{R}^n}}{\operatorname{argmin}} \frac{1}{2} ||Y - X\beta - \gamma||^2 + \rho ||\gamma||_1, \tag{5}$$

where $Y = (Y_1, ..., Y_n) \in \mathbb{R}^n$ are the responses, $X \in \mathbb{R}^{n \times p}$ is the design matrix, and ρ is a penalty hyperparameter to be tuned. Though the motivation for the ℓ_1 penalty from high-dimensional statistics appears somewhat distant from robustness, it turns out [53, 30, 21, 1] that the $\hat{\beta}$ solving (5) is also a robust M-estimator, namely, it solves

$$\min_{\beta \in \mathbb{R}^p} \sum_{j=1}^n \mathcal{L}_{\text{Huber}}(Y_j - \langle x_j, \beta \rangle; \rho),$$

where $\mathcal{L}_{\text{Huber}}$ is the Huber loss function

$$\mathcal{L}_{\text{Huber}}(t; \rho) = \begin{cases} \frac{t^2}{2} & \text{if } |t| \le \rho, \\ \rho|t| - \frac{\rho^2}{2} & \text{if } |t| > \rho. \end{cases}$$

Owen and She [55], motivated from $\hat{\beta}$'s sensitivity to leverage points, suggest modifying (5). Since the ℓ_1 penalty in (5) corresponds to soft-thresholding in $\hat{\gamma}$, they suggest replacing soft-thresholding by other thresholding functions, even those which may correspond to nonconvex penalties. Their article is methodological in nature and no theoretical guarantees are offered. In more modern contexts, the regression parameter β is also sparse. Consequently, a number of articles [47, 18, 16] (see also [29] and references therein for a compressed sensing perspective) investigate a natural extension of (5) by placing an additional ℓ_1 penalty on β.

Returning to our mean shift location model (1), Collier and Dalalyan's article [16] perhaps bears down the most among the aforementioned literature. They study the estimator

$$(\hat{\theta}, \hat{\gamma}) = \underset{\substack{\mu \in \mathbb{R} \\ \gamma \in \mathbb{R}^n}}{\operatorname{argmin}} ||X - \mu \mathbf{1}_n - \gamma||^2 + \rho ||\gamma||_1,$$

where $X = (X_1, ..., X_n) \in \mathbb{R}^n$. As Collier and Dalalyan note, the solution $\hat{\theta}$ has the interpretation of maximizing a profile penalized likelihood

$$\hat{\gamma} = \underset{\gamma \in \mathbb{R}^n}{\operatorname{argmin}} ||(X - \bar{X}\mathbf{1}_n) - (\gamma - \bar{\gamma}\mathbf{1}_n)||^2 + \rho ||\gamma||_1,$$

$$\hat{\theta} = \frac{1}{n} \sum_{j=1}^n (X_j - \hat{\gamma}_j).$$

Assuming $\frac{k}{n} \le c$ for a sufficiently small universal constant $0 < c < \frac{1}{2}$ and $\sigma = 1$, they show that for the choice $\rho \asymp \sqrt{\log n}$ the estimator achieves $|\hat{\theta} - \theta|^2 \lesssim \frac{1}{n} + \frac{k^2}{n^2} \log n$ with high probability. This is actually slower than the rate $\frac{1}{n} + \frac{k^2}{n^2}$ achievable by sample median! The logarithmic factor is a consequence of a typical analysis for ℓ_1-penalized estimators. Collier and Dalalyan suggest a modification by introducing individual penalties, defining $\hat{\gamma} = \operatorname{argmin}_{\gamma \in \mathbb{R}^n} ||(X - \bar{X}\mathbf{1}_n) - (\gamma - \bar{\gamma}\mathbf{1}_n)||^2 + \sum_{j=1}^n \rho_j |\gamma_j|$ and $\hat{\theta} = \frac{1}{n} \sum_{j=1}^n (X_j - \hat{\gamma}_j)$. An optimal choice of the ρ_j depends on unknown quantities. To circumvent this issue, they suggest an iterative scheme and propose an iterative soft-thresholding estimator $\hat{\theta}_{\text{IST}}$. Under the assumption $\frac{k}{n} \le c$, they show $|\hat{\theta}_{\text{IST}} - \theta|^2 \lesssim \frac{1}{n} + \frac{k^2}{n^2}$ with high probability. Hence, $\hat{\theta}_{\text{IST}}$ matches the performance of sample median.

These results may give the impression that the Huber rate $\frac{1}{n} + \frac{k^2}{n^2}$, which is minimax optimal under (2) or (3), should also be the sharp rate in (1). This impression is strengthened when one

3

expresses the observations as

$$X_j \overset{ind}{\sim} \begin{cases} N(\theta, \sigma^2) & if\ j \in \mathcal{I}, \\ N(\eta_j, \sigma^2) & if\ j \in \mathcal{O}, \end{cases} \tag{6}$$

where $\mathcal{O} = \{1 \le j \le n : \gamma_j \ne 0\}$, $\mathcal{I} = \mathcal{O}^c$, and $\eta_j = \theta + \gamma_j$. The expression (6) bears a striking resemblance to (2). The temptation is strong to intuit that since the shifts γ_j in (2) are completely arbitrary as in (6), estimation should not be easier than in the Huber model. In other words, intuition may suggest the Gaussian character of the outliers in (6) is an inconsequential difference.

Though easily forgiven, such intuition is mistaken. Consider the corresponding Bayes formulation (i.e. analogous to how (3) corresponds to (2)),

$$X_1, ..., X_n \overset{iid}{\sim} (1 - \varepsilon)N(\theta, \sigma^2) + \varepsilon(Q * N(0, \sigma^2)), \tag{7}$$

where $\varepsilon = \frac{k}{n}$ and $*$ denotes convolution. Minimax rate conclusions are the same in (7) and (6) (in the way those in (3) and (2) are the same as discussed earlier). A comparison between (7) and the Bayes formulation of Huber's contamination model (3) reveals the additional structure available in the mean shift model. To elaborate, in (3) the distribution with weight ε can be any distribution Q. In contrast, the distribution with weight ε in (7) must be of the form $Q * N(0, \sigma^2)$. Clearly the set of distributions of this form is a smaller set than the set of all distributions.

It is immediately seen that if a rate-optimal estimator is to achieve a faster rate than that in Huber's model, it *must* make specific use of the Gaussian character of the data. The penalized regression estimators discussed earlier appear not to exploit this structure. A fundamental question arises; for which values of k is consistent estimation of θ possible, and for which values is it impossible? More specifically, is the regime of consistency different from that in Huber's model?

1.2 Large-scale inference

In many modern scientific fields, the researcher is faced with the problem of simultaneously considering a very large number of hypothesis tests. For the sake of discussion, let us consider a researcher armed with z-scores $X_1, ..., X_n$, each of which represents a test statistic corresponding to a hypothesis test. Traditionally, it is assumed a z-score follows exactly the standard normal distribution if its corresponding hypothesis is null. To simplify matters, let us also assume the z-score follows a normal distribution with unit variance but nonzero mean if its corresponding hypothesis is nonnull, that is to say,

$$X_j \sim \begin{cases} N(0, 1) & if\ j \in \mathcal{H}_0, \\ N(\gamma_j, 1) & if\ j \in \mathcal{H}_0^c, \end{cases}$$

where $\mathcal{H}_0$ denotes the set of null hypotheses and the nonzero γ_j are the nonnull effects. Assume $|\mathcal{H}_0^c| \le k < \frac{n}{2}$, which is usual in the relevant scientific settings.

Efron, in a sequence of highly original and persuasive articles [23, 24, 25], demonstrated the assumption that the null z-scores exactly follow the standard normal distribution is not at all mild. Seemingly slight misspecification can substantially affect downstream inferential conclusions. As Efron points out, a variety of practical issues can cause misspecification of the null hypothesis.

One major issue is correlation between z-scores [24, 49, 52, 51]. To illustrate in a stylized benchmark example, consider a one factor correlation model [28, 27, 41, 40], where the z-scores are given by $X_j = \gamma_j + \sqrt{\rho}W + \sqrt{1 - \rho}Z_j$ with $W, Z_1, ..., Z_n \overset{iid}{\sim} N(0, 1)$. Here, W is the common factor inducing equicorrelation of level ρ. Observe that, marginally, the null distribution is correctly specified as $X_j \sim N(0, 1)$ if $j \in \mathcal{H}_0$. However, due to the correlation, the ensemble of null z-scores does not behave like a standard normal distribution. Conditionally on W,

$$X_j | W \overset{ind}{\sim} \begin{cases} N(\sqrt{\rho}W, 1 - \rho) & if\ j \in \mathcal{H}_0, \\ N(\sqrt{\rho}W + \gamma_j, 1 - \rho) & if\ j \in \mathcal{H}_0^c. \end{cases}$$

It is more appropriate to treat the ensemble of null z-scores as being drawn from a shifted normal distribution with variance $1 - \rho$, rather than the centered normal distribution with unit variance. As Efron [24, 25] illustrates with more sophisticated arguments, correlation can substantially shift and alter the width of the ensemble null z-scores, and subsequently can have disastrous effects on downstream inference if a standard normal null distribution is assumed (see also [52, 28]). Correlation's effects and methodology to handle it were further studied in [54, 2, 57, 56, 31, 48], but a complete theoretical picture has not yet been attained.

To address issues stemming from a misspecified *theoretical null*, Efron argues an *empirical null* capturing the null z-score ensemble should be estimated from the data. The large-scale nature of the problem is now to our advantage as the majority of the z-scores are nulls; the nonnull z-scores are a nuisance from this point of view. Model (1) is the result of taking Efron's recommendation seriously. Retaining Gaussianity as in [23], Efron suggests considering

$$X_j \overset{ind}{\sim} \begin{cases} N(\theta, \sigma^2) & \text{if } j \in \mathcal{H}_0, \\ N(\eta_j, \sigma^2) & \text{if } j \in \mathcal{H}_0^c, \end{cases} \tag{8}$$

where $\theta \in \mathbb{R}$ and $\sigma > 0$ are parameters to be estimated. The models (1) and (8) are equivalent with $\eta_j = \theta + \gamma_j$ and $\mathcal{H}_0 = \{1 \leq j \leq n : \gamma_j = 0\}$. As in our discussion of robust statistics, the model (8) is closely related to a mixture formulation (7), which has the name *(Gaussian) two-groups model* in the context of large-scale inference. Indeed, minimax conclusions in the two models are the same. Further, Q now has the interpretation of a signal distribution rather than contamination. The two-groups model has been widely adopted for large-scale inference and tremendous methodological development has occurred [56, 36, 7, 58, 8].

It is worth emphasizing the setting $k \asymp n$ is especially relevant to modern large-scale inference. Indeed, it is not at all an extreme case that a constant fraction (e.g. $0.1\%, 1\%, 10\%, 25\%$ etc.) of the z-scores are nonnull. The limitations of the robust statistics literature are readily apparent as the best known estimators for θ are not even consistent in this regime! The minimax rate in Huber's contamination model $\frac{1}{n} + \frac{k^2}{n^2} \asymp 1$ is of constant order when $k \asymp n$. Even in the mean shift setting, the best estimators in the literature are only shown to achieve Huber's rate, to the best of our knowledge. Moreover, special care should be taken to avoid making assumptions like $\frac{k}{n} \leq c$ for some sufficiently small constant c, even though such assumptions seem innocuous in robust statistics. Theory relevant to the practice of large-scale inference ought to address the case where k is quite close to $\frac{n}{2}$. The question raised earlier concerning the possibility of consistent estimation is a pressing one.

In the large-scale inference literature, arguably the most successful estimators for the empirical null parameters in the contexts of (6) and (7) are based on Fourier analysis. To the best of our knowledge, this suite of methods was first introduced and developed in [36, 7, 35]. Fourier-analytic approaches have also seen success in other statistical topics such as deconvolution and estimation in stochastic processes [46, 5, 4]. Also, a Fourier-based estimator was recently shown to nearly achieve the optimal estimation rate of a robust location estimation problem in a heavy-tailed formulation of robustness [3]. Most directly related to our setting is the work of Cai and Jin [7]. To roughly describe the estimators of Cai and Jin [7], observe the characteristic function of (7) is

$$\psi(\omega) := e^{-\frac{\omega^2 \sigma^2}{2}} \left((1 - \varepsilon) e^{i\omega\theta} + \varepsilon \hat{Q}(\omega) \right),$$

where $\hat{Q}$ denotes the characteristic function of Q. Note that $\hat{Q}$ can also be considered in the frequentist context of (8) where now $\hat{Q}(\omega) = \frac{1}{|\mathcal{O}|} \sum_{j \in \mathcal{O}} e^{i\omega\eta_j}$. Cai and Jin's key observation is that for large frequencies ω,

$$\psi(\omega) \approx e^{-\frac{\omega^2 \sigma^2}{2}} (1 - \varepsilon) e^{i\omega\theta},$$

if it is assumed $\hat{Q}$ decays fast enough (e.g. if Q is smooth enough). Hence, the null parameters can be extracted from the large frequency components of the characteristic function. Of course, the population characteristic function is not accessible, so it is estimated via the empirical characteristic

function. Taking an appropriate ω to balance a typical bias-variance tradeoff, estimators $\hat{\theta}_{\mathrm{CJ}}$ and $\hat{\sigma}^2_{\mathrm{CJ}}$ are formed.

Their result can be stated informally as follows. Fix $\alpha > 2$ and $\beta \in [0, 1/2)$. Let $cn^{-\beta} \leq \varepsilon \leq c$ where $0 < c < 1$ is a universal constant. Note c can be larger than $\frac{1}{2}$. Assuming some moment conditions on Q as well as the decay conditions $\limsup_{\omega \to \infty} |\hat{Q}(\omega)||\omega|^{\alpha} < \infty$ and $\limsup_{\omega \to \infty} |\hat{Q}'(\omega)||\omega|^{\alpha+1} < \infty$, they establish

$$E\left(|\hat{\theta}_{\mathrm{CJ}} - \theta|^2\right) \lesssim \frac{\varepsilon^2}{\log^{\alpha+1}(n)}, \tag{9}$$

$$E\left(|\hat{\sigma}^2_{\mathrm{CJ}} - \sigma^2|^2\right) \lesssim \frac{\varepsilon^2}{\log^{\alpha+2}(n)}. \tag{10}$$

Matching minimax lower bounds are also obtained. From these bounds, Cai and Jin [7] have shown that consistent estimation is possible even in the practically salient case $\varepsilon \asymp 1$. Furthermore, their result is quite appealing as it sharply captures the effect of the interaction between the smoothness of the signal distribution Q and the signal sparsity on optimal estimation rates.

Though impressive, their results leave a few fundamental questions unresolved. First, they require a number of assumptions on the signal distribution Q which are undesirable from a robust statistics perspective; nothing can be immediately said about (7) or (1). Noting $\hat{Q}(\omega) = \frac{1}{|\mathcal{O}|} \sum_{j \in \mathcal{O}} e^{i\omega\eta_j}$, the condition $\limsup_{\omega \to \infty} |\hat{Q}'(\omega)||\omega|^{\alpha+1} < \infty$ requires $\limsup_{\omega \to \infty} \left|\frac{1}{|\mathcal{O}|} \sum_{j \in \mathcal{O}} i\eta_j e^{i\omega\eta_j}\right| |\omega|^{\alpha+1} < \infty$. The outliers η_j are effectively required to be bounded. Second, their minimax theory requires the null parameters θ and σ^2 to be bounded by fixed constants. Third, one might hope that plugging in $\alpha = 0$ may give the correct rates of the problem without smoothness. This is not true. The bound (9) applies for any $cn^{-\beta} \leq \varepsilon \leq c$, that is, it notably holds for the choice $\varepsilon = \frac{1}{2}$. Plugging in $\alpha = 0$ gives the rate $\log^{-1}(n)$, asserting that consistent estimation is possible in (7) even when $\varepsilon = \frac{1}{2}$, which is clearly false since θ is not identifiable when $\varepsilon = \frac{1}{2}$ without any assumptions on Q. Hence, the problem of establishing the optimal estimation rate without any assumptions on Q is still completely open.

1.3 Main contribution

Throughout the remainder of the paper, we focus on the frequentist model (1) though we remind the reader all our results also hold for the Bayes formulation (7). Recall $1 \leq k < \frac{n}{2}$ for identifiability reasons. Our first contribution is to obtain the sharp minimax rate of estimating θ with respect to square loss,

$$\epsilon^*(k, n)^2 \asymp \begin{cases} \frac{\sigma^2}{n} & \text{if } 1 \leq k \leq \sqrt{n}, \\ \frac{\sigma^2 k^2}{n^2} \log^{-1}\left(\frac{ek^2}{n}\right) & \text{if } \sqrt{n} < k \leq \frac{n}{4}, \\ \sigma^2 \log^{-1}\left(\frac{e(n-2k)^2}{n}\right) & \text{if } \frac{n}{4} < k \leq \frac{n}{2} - \sqrt{n}, \\ \sigma^2 \log\left(\frac{en}{(n-2k)^2}\right) & \text{if } \frac{n}{2} - \sqrt{n} < k < \frac{n}{2}. \end{cases} \tag{11}$$

The minimax rate turns out to exhibit scaling not only in k but also in $n - 2k$. Scaling of the form $n - k$ has been observed in a sparse signal detection problem under correlation [37] as well as a one-sided version of (1) where the shape constraints $\gamma_i \geq 0$ are imposed [9]. It is quite natural to see $n - k$ as it denotes the number of "null" or "uncontaminated" samples. The term $n - 2k$ denotes the number of inliers (nulls) minus the number of outliers (nonnulls). Its appearance is not surprising since the problem becomes more difficult as k approaches $\frac{n}{2}$, at which point θ is no longer identifiable.

The fundamental question regarding when consistent estimation is possible can now be answered. It is quickly concluded that consistent estimation of θ is possible if and only if $n - 2k = \omega(\sqrt{n})$. In contrast, consistent estimation is possible in Huber's model (2) if and only if $k = o(n)$. Importantly, θ can be consistently estimated in (1) in the practically relevant setting where k is a constant fraction of n mentioned in Section 1.2.

When $n - 2k \lesssim \sqrt{n}$, consistent estimation of θ is not possible. The minimax rate in this regime

can be directly established from recently available results in [38]. A kernel mode estimator turns out to be rate-optimal. Section 5.1 discusses this regime further.

We note that the minimax rate in the regime $1 \leq k \leq cn$ for a small universal constant $c \in (0, 1)$ had been independently developed by Carpentier and Verzelen[1]. Though their result nicely pins down the rate in this regime, the remaining regime is unaddressed. Our result delivers a full characterization of the minimax rate by covering the practically important case where k can be a large fraction of n. Broadly, both our upper and lower bound approaches are conceptually different from theirs.

Remark 1 (The effect of the data's Gaussian character). By comparing to Huber's contamination model (2), our result enables us to understand the effect of the data's Gaussian character. To elaborate, in both (1) and Huber's contamination model (2) the minimax rate is parametric for $k \leq \sqrt{n}$. For $\sqrt{n} \leq k \leq \frac{n}{5}$ (recall [13] assumes the contamination level is below 1/5), there is a $\log^{-1}(k^2/n)$ improvement. It turns out one can easily prove (see Appendix E), using the methods of [13], that the minimax rate in Huber's model (2) is $\log(en/(n - 2k))$ for $\frac{n}{5} \leq k < \frac{n}{2}$. Hence, when $n - 2k = o(n)$ and $n - 2k = \omega(\sqrt{n})$ the rate in Huber's model is actually growing whereas $\epsilon^*(k, n)^2 \asymp \log^{-1}\left(e(n - 2k)^2/n\right)$ is vanishing. All of these improvements are attributable solely to the Gaussian character of the data.

Remark 2 (Comparison to a one-sided version). It is also interesting to compare the location estimation rate $\epsilon^*(k, n)^2$ to the minimax rate obtained for the one-sided version of (1) in [9] where $\gamma_i \geq 0$,

$$\epsilon^*_{\text{one-sided}}(k, n)^2 \asymp \begin{cases} \frac{\sigma^2}{n} & \text{if } 1 \leq k \leq \sqrt{n}, \\ \frac{\sigma^2 k^2}{n^2} \log^{-3}\left(\frac{k^2}{n}\right) & \text{if } \sqrt{n} < k < \frac{n}{2}, \\ \frac{\sigma^2 \log^4\left(\frac{n}{n-k}\right)}{\log^3 n} & \text{if } \frac{n}{2} \leq k \leq n - 1. \end{cases}$$

Note k can be as large as $n - 1$ since the shape constraint on γ ensures identifiability of θ. The situation in the one-sided case is quite different from the two-sided case we consider; the rate $\epsilon^*(k, n)$ is not at all an obvious extension of $\epsilon^*_{\text{one-sided}}(k, n)$. The estimators proposed in [9] heavily rely on the shape constraint and thus are not suitable for (1).

Our second contribution is to establish the minimax rate for estimating the variance σ^2,

$$\epsilon^*_{\text{var}}(k, n)^2 \asymp \begin{cases} \frac{1}{n} & \text{if } 1 \leq k \leq \sqrt{n}, \\ \frac{k^2}{n^2} \log^{-2}\left(1 + \frac{k}{\sqrt{n}}\right) & \text{if } \sqrt{n} < k < \frac{n}{2}. \end{cases} \tag{12}$$

This rate should be compared to the minimax rate for variance estimation in Huber's contamination model (3). The article [13] showed, under the assumption $\frac{k}{n} \leq \frac{1}{5}$, the rate $\frac{1}{n} + \frac{k^2}{n^2}$. Namely, it is exactly the same rate for estimating the location parameter θ. In contrast, we see $\epsilon^*_{\text{var}}(k, n)^2$ is not only faster than the rate in Huber's model, but it also enjoys a logarithmic improvement over $\epsilon^*(k, n)^2$.

It also turns out that the minimax rate for variance estimation is exactly the same as if θ were already known. In particular, our result is an extension of the minimax rate for variance estimation in the sparse Gaussian sequence model [17] (i.e. $\theta = 0$ is known). However, the estimator proposed in [17] requires $\frac{k}{n} \leq c$ for a small universal constant $0 < c < \frac{1}{8}$. As discussed in Section 1.2, such a condition is undesirable. The estimator we propose establishes that the minimax rate proved in [17] extends beyond this regime. Variance estimation was also considered in [7], but their result suffers from requiring various assumptions on the outliers which have been discussed in Section 1.2.

As a consequence of our results on parameter estimation, the minimax rate for estimating the null distribution in total variation follows directly. The rate for location estimation turns out to dominate, and so most of the conclusions for estimating θ carry over to minimax estimation in total variation.

A few high-level remarks about our approach are in order. For both location and variance estimation, we take inspiration from earlier work [17, 7, 36, 35] and adopt a Fourier-based approach in both the upper and lower bounds. In Section 2, the location estimator's development is discussed in detail

---

[1](A. Carpentier and N. Verzelen, personal communication, 2023)

for the case of known variance. In Section 3, the estimator is generalized to handle unknown variance; rate-optimal variance estimation is also discussed. For location estimation, the lower bound is proved via a two-point testing argument by constructing two marginal distributions in the Bayes model (7) which have characteristic functions which agree on a large interval containing zero. The construction is notably different from existing work [17, 7] in order to sharply capture the second order $n - 2k$ scaling in the minimax rate; previous constructions miss this scaling entirely. As mentioned earlier, the lower bound for variance estimation was known [17], but a matching upper bound for all $k < \frac{n}{2}$ was absent from the literature.

Our Fourier-based upper bound for location estimation, though similar in examining frequency space, is very different from the estimator proposed by Cai and Jin [7]. A detailed discussion is given in Section 2. Briefly, as noted in Section 1.2, Cai and Jin localize θ by inspecting the empirical characteristic function at a suitably large frequency; the assumptions they place on the outliers are crucial to their estimator's success. A big advantage of the characteristic function is that it transforms the unbounded quantity X_j to the bounded quantity $e^{i\omega X_j}$. Specifically, it is very well-suited to dealing with unbounded outliers. However, the assumptions (e.g. bounded moments of outliers, bounded null parameters) of [7] limit this advantage of the characteristic function. In contrast to [7], we make no assumptions on the outliers and instead consider an estimator which fits to the nuisance outliers in frequency space. Interestingly, the interval of frequencies considered in the upper bound is the same interval of frequencies used in the lower bound, thus pointing to the optimality of the estimator.

1.4 Notation

For $a, b \in \mathbb{R}$ the notation $a \lesssim b$ denotes the existence of a universal constant $c > 0$ such that $a \leq cb$. The notation $a \gtrsim b$ is used to denote $b \lesssim a$. Additionally $a \asymp b$ denotes $a \lesssim b$ and $a \gtrsim b$. The symbol $:=$ is frequently used when defining a quantity or object. Furthermore, we frequently use $a \vee b := \max(a, b)$ and $a \wedge b := \min(a, b)$. We generically use the notation $\mathbb{1}_A$ to denote the indicator function for an event A. For a vector $v \in \mathbb{R}^n$ and a subset $S \subset [n]$, we sometimes use the notation $v_S \in \mathbb{R}^n$ to denote the vector with coordinate j equal to v_j if $j \in S$ and zero otherwise. Additionally, $||v||_0 := \sum_{j=1}^n \mathbb{1}_{\{v_j \neq 0\}}$, $||v||_1 := \sum_{j=1}^n |v_j|$, and $||v||^2 := \sum_{j=1}^n v_j^2$. For two probability measures P and Q on a measurable space $(\mathcal{X}, \mathcal{A})$, the total variation distance is defined as $\mathrm{d}_{\mathrm{TV}}(P, Q) := \sup_{A \in \mathcal{A}} |P(A) - Q(A)|$. If P is absolutely continuous with respect to Q, then the χ^2-divergence is defined as $\chi^2(P||Q) := \int_{\mathcal{X}} \left(\frac{dP}{dQ} - 1 \right)^2 dQ$. We will frequently use the same notation for two probability densities p and q. For sequences $\{a_k\}_{k=1}^\infty$ and $\{b_k\}_{k=1}^\infty$, the notation $a_k = o(b_k)$ denotes $\lim_{k \to \infty} \frac{a_k}{b_k} = 0$ and the notation $a_k = \omega(b_k)$ is used to denote $b_k = o(a_k)$. For a point $x \in \mathbb{R}$, the symbol δ_x denotes the probability measure which places full probability mass at the point x. The symbol $*$ denotes convolution and the same symbol will be used in the context of the convolution of probability measures as well as functions. Complex numbers are used throughout and so $i \in \mathbb{C}$ denotes the imaginary unit satisfying $i^2 = -1$. For $z \in \mathbb{C}$, $\bar{z}$ denotes the complex conjugate of z. In other places, i may be used as an indexing variable; the context will be clear to the reader even if not explicitly stated. Furthermore, we will $|\cdot|$ to denote modulus. In some places, it is used with a finite set, in which case it denotes cardinality.

2 A Fourier-based estimator

In this section, we propose a Fourier-based location estimator for the regime $k \leq \frac{n}{2} - C\sqrt{n}$. According to the minimax rate formula (11), this is the regime where consistent estimation is possible. A rate-optimal estimator in the other regime $k > \frac{n}{2} - C\sqrt{n}$ will be presented in Section 5.1.1. For ease of presentation and to aid understanding, we first discuss the case where the variance is known and taken to be unit. For notational ease in this section, $P_{\theta,\gamma}$ and $E_{\theta,\gamma}$ will be used in place of $P_{\theta,\gamma,1}$ and $E_{\theta,\gamma,1}$ respectively.

Inspired by the success of Fourier-based methods [36, 7], we look to the empirical characteristic function. As discussed in Section 1.2, the central idea of [7, 36] is no longer applicable since the characteristic function of Q need not decay when no assumptions are imposed. Before defining our estimator, we briefly outline Cai and Jin's estimator [7]. Their article works in the context of the Bayes model (7) but is also applicable in the frequentist model (6) by taking $Q = \frac{1}{|\mathcal{O}|} \sum_{j \in \mathcal{O}} \delta_{\eta_j}$. Let $\psi(\omega) = e^{-\frac{\omega^2 \sigma^2}{2}} \left((1 - \varepsilon)e^{i\omega\theta} + \varepsilon \hat{Q}(\omega) \right)$ and $\psi_0(\omega) = e^{i\omega\theta - \frac{\omega^2 \sigma^2}{2}}$ denote the characteristic functions of the marginal and the null distributions. To extract θ from the characteristic function, Cai and Jin [7, 36] first define the function $\mu : \mathbb{R} \to \mathbb{R}$ with

$$\mu(\omega; \xi) = \left. \frac{\Im(\overline{\xi(t)}\xi'(t))}{|\xi(t)|^2} \right|_{t=\omega}$$

for $\omega \in \mathbb{R}$ and for a differentiable function $\xi : \mathbb{R} \to \mathbb{C}$. Here, $\Im(z)$ denotes the imaginary part of the complex number $z \in \mathbb{C}$. Importantly, it is shown that $\mu(\omega; \psi_0) = \theta$ for all $\omega \neq 0$. Of course, ψ_0 is not available to the statistician. Taking the empirical characteristic function $\hat{\psi}(\omega) := \frac{1}{n} \sum_{j=1}^{n} e^{i\omega X_j}$, Cai and Jin [7] define the estimator $\hat{\theta} = \mu(\omega^*; \hat{\psi})$ with a specific choice of ω^*. As explained in [7, 36], the idea is that plugging in the empirical characteristic function should yield $\hat{\theta} = \mu(\omega^*, \hat{\psi}) \approx \mu(\omega^*, \psi)$. Since $\theta = \mu(\omega^*, \psi_0)$, Cai and Jin [7] essentially need ψ to be close to ψ_0. Since $\psi(\omega) = (1 - \varepsilon)\psi_0(\omega) + \varepsilon \hat{Q}(\omega)e^{-\frac{\omega^2\sigma^2}{2}}$, assumptions on Q cannot be avoided if ψ and ψ_0 are to be close. This is a major drawback to writing θ as a function of ψ_0.

We obtain θ from ψ directly and thus do not need ψ and ψ_0 to be close. Consequently, assumptions on the outliers are totally avoided. To motivate our construction, consider the following line of reasoning. Recall the notation $\mathcal{O} = \{1 \leq j \leq n : \gamma_j \neq 0\}$ and $\mathcal{I} = \mathcal{O}^c$. Consider the expectation of the empirical characteristic function at any frequency $\omega \in \mathbb{R}$ is

$$E_{\theta,\gamma}\left(\frac{1}{n} \sum_{j=1}^{n} e^{i\omega X_j} \right) = e^{i\omega\theta - \frac{\omega^2}{2}} \left(\frac{n - |\mathcal{O}|}{n} + \frac{1}{n} \sum_{j \in \mathcal{O}} e^{i\omega\gamma_j} \right).$$

Cai and Jin [7] essentially assume $\frac{1}{|\mathcal{O}|} \sum_{j \in \mathcal{O}} e^{i\omega\gamma_j}$ is small for a choice of large frequency. Instead, we will fit the function $\omega \mapsto \frac{1}{|\mathcal{O}|} \sum_{j \in \mathcal{O}} e^{i\omega\gamma_j}$ directly. The following result states how θ can be obtained from a population level optimization program.

Lemma 2.1. *If $k < \frac{n}{2}$, then*

$$\theta = \operatorname*{argmin}_{\mu \in \mathbb{R}} \sup_{\omega \in \mathbb{R}} \inf_{\substack{\zeta \in \mathbb{C} \\ |\zeta| \leq 1}} \left| E_{\theta,\gamma}\left(\frac{1}{n} \sum_{j=1}^{n} e^{i\omega(X_j - \mu) + \frac{\omega^2}{2}} \right) - \frac{n - k}{n} - \frac{k}{n}\zeta \right|. \tag{13}$$

Proof. Clearly θ is a minimizer at which the objective function is equal to zero. It remains to show θ is a unique minimizer. Suppose μ is also a minimizer. Then

$$0 = \sup_{\omega \in \mathbb{R}} \inf_{\substack{\zeta \in \mathbb{C} \\ |\zeta| \leq 1}} \left| E_{\theta,\gamma}\left(\frac{1}{n} \sum_{j=1}^{n} e^{i\omega(X_j - \mu) + \frac{\omega^2}{2}} \right) - \frac{n - k}{n} - \frac{k}{n}\zeta \right|$$

$$= \sup_{\omega \in \mathbb{R}} \inf_{\substack{\zeta \in \mathbb{C} \\ |\zeta| \leq 1}} \left| \frac{n - k}{n} \left(e^{i\omega(\theta - \mu)} - 1 \right) + \frac{k}{n} \left(\frac{1}{k} \sum_{j \in \mathcal{O}'} e^{i\omega(\theta + \gamma_j - \mu)} - \zeta \right) \right|$$

where $\mathcal{O}'$ is $\mathcal{O}$ along with some arbitrary indices taken from $\mathcal{I}$ (if needed) to ensure $|\mathcal{O}'| = k$. Suppose

$\mu \neq \theta$. By reverse triangle inequality,

$$
\begin{aligned}
0 &\geq \sup_{\omega \in \mathbb{R}} \inf_{\substack{\zeta \in \mathbb{C} \\ |\zeta| \leq 1}} \left| \frac{n-k}{n} \left(e^{i\omega(\theta-\mu)} - 1 \right) \right| - \left| \frac{k}{n} \left(\frac{1}{k} \sum_{j \in \mathcal{O}'} e^{i\omega(\theta+\gamma_j-\mu)} - \zeta \right) \right| \\
&\geq \sup_{\omega \in \mathbb{R}} \frac{n-k}{n} \left| e^{i\omega(\theta-\mu)} - 1 \right| - \frac{2k}{n} \\
&\geq \frac{n-k}{n} \left| e^{i\frac{\pi}{\theta-\mu}(\theta-\mu)} - 1 \right| - \frac{2k}{n} \\
&= \frac{2(n-2k)}{n}
\end{aligned}
$$

which is a contradiction since $k < \frac{n}{2}$. Therefore, $\theta = \mu$, which is to say θ is the unique minimizer. $\quad\square$

Since θ can be written as the minimizer of a population level optimization program, it is natural to consider the estimator which minimizes an empirical version of the population level program. Following instinct, we would like to replace the expectation in (13) with its empirical counterpart. However, it is nonsensical to do this without any modification since (13) involves supremum over $\omega \in \mathbb{R}$ and so the term $\frac{1}{n} \sum_{j=1}^{n} e^{i\omega(X_j-\mu)+\frac{\omega^2}{2}}$ has variance which may blow up. To rectify this issue, we truncate and only consider $|\omega| \leq \tau$. Consider

$$
\hat{\theta} = \operatorname*{argmin}_{\mu \in \mathbb{R}} \sup_{|\omega| \leq \tau} \inf_{\substack{\zeta \in \mathbb{C} \\ |\zeta| \leq 1}} \left| \frac{1}{n} \sum_{j=1}^{n} e^{i\omega(X_j-\mu)+\frac{\omega^2}{2}} - \frac{n-k}{n} - \frac{k}{n}\zeta \right|. \tag{14}
$$

Here, τ is a tuning parameter to be selected. It will turn out to determine the bias-variance tradeoff in the risk of $\hat{\theta}$.

Theorem 2.2. *Fix $\delta \in (0,1)$. There exist $C, C', c > 0$ depending only on δ such that if n is sufficiently large depending only on δ, $k \leq \frac{n}{2} - C\sqrt{n}$, and $\tau = 1 \vee c\sqrt{\log\left(1 + \frac{k^2(n-2k)^2}{n^3}\right)}$, then*

$$
\sup_{\substack{\theta \in \mathbb{R} \\ ||\gamma||_0 \leq k}} P_{\theta,\gamma}\left\{ |\hat{\theta} - \theta| > C'\left(\frac{1}{\sqrt{n}} + \frac{k}{n}\tau^{-1} \right) \right\} \leq \delta.
$$

Note

$$
\frac{1}{\sqrt{n}} + \frac{k}{n\sqrt{1 \vee \log\left(1 + \frac{k^2(n-2k)^2}{n^3}\right)}} \asymp \begin{cases} \frac{1}{\sqrt{n}} & \text{if } k \leq \sqrt{n}, \\ \frac{k}{n\sqrt{\log\left(1 + \frac{k^2}{n}\right)}} & \text{if } \sqrt{n} < k < \frac{n}{4}, \\ \frac{k}{n\sqrt{\log\left(1 + \frac{(n-2k)^2}{n}\right)}} & \text{if } \frac{n}{4} \leq k < \frac{n}{2} - \sqrt{n}, \end{cases}
$$

from the inequality $u/2 \leq \log(1+u) \leq u$ for $u \in (0,1)$. Thus, $\hat{\theta}$ indeed achieves the rate (11) when $k \leq \frac{n}{2} - C\sqrt{n}$.

As with other Fourier-based methods, $\hat{\theta}$ exploits the Gaussian character of the data to denoise. Namely, (14) involves dividing by the characteristic function of the standard Gaussian distribution, i.e. the term $e^{\omega^2/2}$ appears. As discussed in Section 5.4, the estimator can be generalized to handle non-Gaussian noise. A couple of remarks are in order.

Remark 3. The minimax rate for $1 \leq k \leq cn$ for a small universal constant $c \in (0,1)$ had been independently developed by Carpentier and Verzelen[2], though their estimator is conceptually different from ours.

Remark 4 (Computation). Fitting to the outliers in the frequency domain is advantageous as it only involves optimization over the two-dimensional (one-dimensional complex) variable ζ. In contrast, typical ideas of fitting to outliers in the spatial domain (e.g. as in the mean shift contamination literature discussed in Section 1.1) require optimization over a k-dimensional variable (or an n-dimensional variable with a penalty, as in [16] mentioned in Section 1.1). The computation of (14) is

---

[2](A. Carpentier and N. Verzelen, personal communication, 2023)

thus straightforward. First, consider the minimization over μ. One can take the interval having length of order $\sqrt{\log n}$ centered at median$(X_1, ..., X_n)$ and discretize it with grid points having distance of order the statistical rate $\epsilon(k, n)$ given by (11). Since the domain of optimization over ω is a bounded interval, it can be discretized with a fine enough grid. A glance at the proof suggests taking grid points having distance of order $\frac{1}{\sqrt{n \log n}}$ since we are guaranteed $|\hat{\theta} - \theta| \lesssim \sqrt{\log n}$ by the discretization employed for μ. Minimization over ζ is direct as it is a convex problem with a convex constraint. The estimator can then be computed by a simple two-dimensional grid search.

3 Methodology: unknown variance

In this section, the model (1) is considered in the setting where the variance σ^2 is unknown. Sections 3.1 and 3.2 address variance estimation and Section 3.3 addresses generalizing the Fourier-based estimator of Section 2 to handle unknown variance.

3.1 A pilot variance estimator

A pilot estimator which captures the order of σ^2 will be needed, and we will directly use the correlation estimator of [38]. As the context of [38] appears different from our current setting, we discuss, for the reader's understanding, the development of the pilot estimator of [38]. One idea for a variance estimator is

$$\min_{|S| = \lceil \frac{n}{2} \rceil} \frac{1}{|S| - 1} \sum_{j \in S} (X_j - \bar{X}_S)^2,$$

where $\bar{X}_S = \frac{1}{|S|} \sum_{j \in S} X_j$. The intuition is that the X_j all have the same mean θ for $j \in \mathcal{I}$, and so the usual sample variance estimator could be used if $\mathcal{I}$ were known. Since it is unknown, a search must be done. Subsets of size at least $\frac{n}{2}$ are searched over because $\mathcal{I}$ is of at least this size. If $S \subset \mathcal{I}$, then $\frac{1}{|S| - 1} \sum_{j \in S} (X_j - \bar{X}_S)^2$ will be a decent estimator of σ^2. On the other hand, if $S \cap \mathcal{O} \neq \emptyset$, then $\frac{1}{|S| - 1} \sum_{j \in S} (X_j - \bar{X}_S)^2$ will overestimate σ^2. Outliers can be conceptualized as contributing additional variability. Therefore, it is natural to minimize over all subsets $S \subset \{1, ..., n\}$ with $|S| = \lceil \frac{n}{2} \rceil$. However, the variance estimator in our setting appears to require an exhaustive search and thus seems computationally intractable. The idea is rescued in [38] by random sampling to obtain a polynomial-time estimator.

Independently draw subsets $E_1, ..., E_m \subset \{1, ..., n\}$ of size ℓ uniformly at random and define

$$\tilde{\sigma}^2 := \min_{1 \leq r \leq m} \frac{1}{\ell - 1} \sum_{j \in E_r} (X_j - \bar{X}_{E_r})^2, \tag{15}$$

where $\bar{X}_{E_r} = \frac{1}{\ell} \sum_{j \in E_r} X_j$. It is immediate that $\tilde{\sigma}^2$ is an estimator computable in polynomial time in n whenever m scales polynomially in n. The following result can be proved via the same argument as in [38], so we omit the proof.

Proposition 3.1 ([38]). *Fix $\delta \in (0, 1)$. There exist constants $C_1, C_2, L > 0$ depending only on δ such that if $m = \lceil n^{C_1} \rceil$ and $2 < \ell = \lceil C_2 \log n \rceil$, then*

$$\inf_{\substack{\theta \in \mathbb{R} \\ ||\gamma||_0 < n/2 \\ \sigma > 0}} P_{\theta, \gamma, \sigma} \left\{ L^{-1} \leq \frac{\tilde{\sigma}^2}{\sigma^2} \leq L \right\} \geq 1 - \delta.$$

With the pilot estimator in hand, a Fourier-based variance estimator can now be constructed.

3.2 A rate-optimal variance estimator

Our Fourier-based variance estimator is quite different from Cai and Jin's Fourier-based variance estimator [7]. Recalling they work in the context of (7) (but their work is applicable to (6) as well), let $\psi(\omega) = e^{-\omega^2 \sigma^2 / 2} \left((1 - \varepsilon) e^{i \omega \theta} + \varepsilon \hat{Q}(\omega) \right)$ and $\psi_0(\omega) = e^{i \omega \theta - \omega^2 \sigma^2 / 2}$ denote the characteristic

functions of the marginal and the null distributions. Cai and Jin [7, 36] define the function $v : \mathbb{R} \to \mathbb{R}$ with

$$v(\omega; \xi) = - \left. \frac{\frac{d}{dt}|\xi(t)|}{t|\xi(t)|} \right|_{t=\omega}$$

for $\omega \in \mathbb{R}$ and for a differentiable function $\xi : \mathbb{R} \to \mathbb{C}$. Importantly, $v(\omega; \psi_0) = \sigma^2$ for all $\omega \neq 0$. With the empirical characteristic function $\hat{\psi}(\omega) := \frac{1}{n} \sum_{j=1}^{n} e^{i\omega X_j}$, Cai and Jin [7] define the estimator $\hat{\sigma}^2 = v(\omega^*; \hat{\psi})$ with some specific choice ω^*. The same considerations noted for their location estimator in Section 2 apply to their variance estimator. As mentioned before, their estimator requires decay of $\hat{Q}$ and so various assumptions were imposed on the outliers.

The estimator we propose avoids assumptions by exploiting the connection between the variance and the norm of the characteristic function. Set $\hat{N}(\omega) := \left| \frac{1}{n} \sum_{j=1}^{n} e^{i\omega X_j} \right|$. Define

$$\hat{\sigma}^2 = \inf_{a \leq \omega \leq b} -\frac{2 \log \hat{N}(\omega)}{\omega^2} \tag{16}$$

where $a \leq b$ are tuning parameters to be set.

Theorem 3.2. *Fix $\delta \in (0, 1)$. There exist constants $C, c > 0$ depending only on δ such that if n is sufficiently large depending only on δ, $a = c\tilde{\sigma}^{-1} \left(\sqrt{1 \vee \log \left(\frac{ek^2}{n} \right)} \right)$, and $b = 100a$, then*

$$\sup_{\substack{\theta \in \mathbb{R} \\ ||\gamma||_0 \leq k \\ \sigma > 0}} P_{\theta, \gamma, \sigma} \left\{ \frac{|\hat{\sigma}^2 - \sigma^2|}{\sigma^2} > \frac{Ck}{n \log \left(1 + \frac{k}{\sqrt{n}} \right)} \right\} \leq \delta,$$

where $\tilde{\sigma}$ is the pilot estimator from Section 3.1.

Note

$$\frac{k}{n \log \left(1 + \frac{k}{\sqrt{n}} \right)} \asymp \begin{cases} \frac{1}{\sqrt{n}} & \text{if } 1 \leq k \leq \sqrt{n}, \\ \frac{k}{n \log \left(1 + \frac{k}{\sqrt{n}} \right)} & \text{if } \sqrt{n} < k < \frac{n}{2} \end{cases}$$

by the inequality $u/2 \leq \log(1 + u) \leq u$ for $u \in (0, 1)$. Thus, $\hat{\sigma}^2$ indeed achieves the rate (12).

The estimator given in (16) is quite close to the variance estimator proposed in the context of sparse Gaussian sequence model [17], which is a special case of (1) with $\theta = 0$. Even though θ is unknown in the setting of (1), it is clear $\hat{N}(\omega)$ is shift invariant; it retains the same value even if the original data $\{X_j\}_{j=1}^{n}$ are replaced with the shifted data $\{X_j + \mu\}_{j=1}^{n}$ for any $\mu \in \mathbb{R}$. The difference between (16) and the estimator of [17] lies entirely in taking infimum over $a \leq \omega \leq b$; the authors of [17] explicitly choose ω and use the estimator

$$\hat{\sigma}_\omega^2 = -\frac{2 \log \hat{N}(\omega)}{\omega^2}. \tag{17}$$

However, it was only proved in [17] that this estimator is rate-optimal in the regime $\frac{k}{n} \leq c$ for some sufficiently small constant $0 < c < \frac{1}{8}$. In large-scale inference contexts, there is a practical need to address the case $k \geq \frac{n}{8}$.

To explain how (16) improves on (17), suppose $|\mathcal{O}| = k$ without loss of generality and consider the population counterpart of $\hat{N}(\omega)$,

$$N(\omega) := \left| \frac{1}{n} \sum_{j=1}^{n} e^{i\omega(\theta + \gamma_j) - \frac{\sigma^2 \omega^2}{2}} \right| = e^{-\frac{\sigma^2 \omega^2}{2}} \left| 1 - \frac{k}{n} \left(1 - \frac{1}{k} \sum_{j \in \mathcal{O}} e^{i\omega \gamma_j} \right) \right|.$$

A rearrangement of the above equality leads to the following population counterpart of (17),

$$-\frac{2 \log N(\omega)}{\omega^2} = \sigma^2 - \frac{2}{\omega^2} \log \left| 1 - \frac{k}{n} \left(1 - \frac{1}{k} \sum_{j \in \mathcal{O}} e^{i\omega \gamma_j} \right) \right|. \tag{18}$$

12

The key quantity in (18) is the nonnegative bias term $-\frac{2}{\omega^2} \log \left| 1 - \frac{k}{n} \left(1 - \frac{1}{k} \sum_{j \in \mathcal{O}} e^{i\omega\gamma_j} \right) \right|$, and the extra infimum over $a \leq \omega \leq b$ in (16) is to achieve a smaller bias. To elaborate, we first note that

$$\left| 1 - \frac{1}{k} \sum_{j \in \mathcal{O}} e^{i\omega\gamma_j} \right|^2 = 1 + \left| \frac{1}{k} \sum_{j \in \mathcal{O}} e^{i\omega\gamma_j} \right|^2 - \frac{2}{k} \sum_{j \in \mathcal{O}} \cos(\omega\gamma_j) \leq 2 \left(1 - \frac{1}{k} \sum_{j \in \mathcal{O}} \cos(\omega\gamma_j) \right). \quad (19)$$

Thus, a lower bound for $\frac{1}{k} \sum_{j \in \mathcal{O}} \cos(\omega\gamma_j)$ leads to an upper bound for the positive bias term in (18). The paper [17] uses the inequality $\frac{1}{k} \sum_{j \in \mathcal{O}} \cos(\omega\gamma_j) \geq -1$, so that

$$-\frac{2}{\omega^2} \log \left| 1 - \frac{k}{n} \left(1 - \frac{1}{k} \sum_{j \in \mathcal{O}} e^{i\omega\gamma_j} \right) \right| \leq \frac{2}{\omega^2} \log \left(\frac{n}{n - 2k} \right). \quad (20)$$

It is clear k needs to be bounded away from $\frac{n}{2}$ so that the bias bound (20) does not blow up. In fact, the bound (20) is sharp in the sense that for any ω, there exist $\gamma_1, \ldots, \gamma_k$ such that (20) is an equality. Therefore, the condition that k is bounded away from $\frac{n}{2}$ imposed by [17] cannot be improved for the estimator (17).

The additional infimum in our estimator (16) leads to better bias control. To elaborate, consider the following population counterpart of (16),

$$\inf_{a \leq \omega \leq b} -\frac{2 \log N(\omega)}{\omega^2} = \sigma^2 - \sup_{a \leq \omega \leq b} \frac{2}{\omega^2} \log \left| 1 - \frac{k}{n} \left(1 - \frac{1}{k} \sum_{j \in \mathcal{O}} e^{i\omega\gamma_j} \right) \right|. \quad (21)$$

Compared with (18), now the positive bias term becomes $-\sup_{a \leq \omega \leq b} \frac{2}{\omega^2} \log \left| 1 - \frac{k}{n} \left(1 - \frac{1}{k} \sum_{j \in \mathcal{O}} e^{i\omega\gamma_j} \right) \right|$. In view of the inequality (19), it suffices to find a lower bound for $\sup_{a \leq \omega \leq b} \frac{1}{k} \sum_{j \in \mathcal{O}} \cos(\omega\gamma_j)$, which is given by the following proposition.

Proposition 3.3. *Suppose* $\gamma_1, \ldots, \gamma_k \in \mathbb{R}$. *Define the function* $f : \mathbb{R} \to \mathbb{R}$ *with* $f(\omega) = \frac{1}{k} \sum_{j=1}^{k} \cos(\omega\gamma_j)$. *If* $\alpha > 0$, *then* $\sup_{\omega \in [\alpha, 100\alpha]} f(\omega) \geq -\frac{1}{5}$.

To apply Proposition 3.3, we set $b = 100a$, and then the bias term in (21) can be bounded by

$$- \sup_{a \leq \omega \leq 100a} \frac{2}{\omega^2} \log \left| 1 - \frac{k}{n} \left(1 - \frac{1}{k} \sum_{j \in \mathcal{O}} e^{i\omega\gamma_j} \right) \right| \leq -\frac{2}{a^2} \log \left(1 - \sqrt{\frac{12}{5}} \frac{k}{n} \right) \asymp \frac{1}{a^2},$$

whenever the condition $\frac{k}{n} < \frac{1}{2}$ holds. In particular, the bias bound does not blow up to infinity when $k = \frac{n}{2}$, and an improvement over (20) for k near $\frac{n}{2}$ has been achieved. The estimator (16) is thus motivated from this bias improvement. The choice of the hyperparameter a involves a typical bias-variance tradeoff and Theorem 3.2 attests the success of (16).

3.3 A variance-adaptive location estimator

In this section, the Fourier-based estimator of Section 2 is generalized to handle unknown variance. Define

$$[\sigma_-^2, \sigma_+^2] := \left[\hat{\sigma}^2 \left(1 - \frac{Rk}{n \log \left(1 + \frac{k}{\sqrt{n}} \right)} \right), \hat{\sigma}^2 \left(1 + \frac{Rk}{n \log \left(1 + \frac{k}{\sqrt{n}} \right)} \right) \right]$$

where R is some constant to be set. Then, the estimator is

$$(\hat{\theta}, \hat{v}) = \underset{\substack{\mu \in \mathbb{R} \\ \sigma_-^2 \leq v \leq \sigma_+^2}}{\operatorname{argmin}} \sup_{|\omega| \leq \tau} \inf_{\substack{\zeta \in \mathbb{C} \\ |\zeta| \leq 1}} \left| \frac{1}{n} \sum_{j=1}^{n} e^{i\omega(X_j - \mu) + \frac{v^2\omega^2}{2}} - \frac{n-k}{n} - \frac{k}{n} \zeta \right|. \quad (22)$$

The following theoretical guarantee is available.

Theorem 3.4. *Fix* $\delta \in (0, 1)$. *Let* $\hat{\sigma}^2$ *be the variance estimator from Section 3.2 at confidence level* $\frac{\delta}{2}$. *There exist* $C, C', c, R > 0$ *depending only on* δ *such that the following holds. If* n *is sufficiently*

large depending only on δ, $1 \le k \le \frac{n}{2} - C\sqrt{n}$, and $\tau = \hat{\sigma}^{-1}\left(1 \vee c\sqrt{\log\left(1 + \frac{k^2(n-2k)^2}{n^3}\right)}\right)$, then

$$\sup_{\substack{\theta \in \mathbb{R} \\ ||\gamma||_0 \le k \\ \sigma > 0}} P_{\theta,\gamma,\sigma}\left\{\frac{|\hat{\theta} - \theta|}{\sigma} > C'\frac{k}{n}\log^{-\frac{1}{2}}\left(1 + \frac{k^2(n-2k)^2}{n^3}\right)\right\} \le \delta$$

where $\hat{\theta}$ is given by (22).

This estimator achieves the same rate as if the variance is known; adaptation to unknown σ^2 is possible. The proof is broadly the same as the argument in the case where σ^2 is known. The only difference is in the additional optimization over v, which is not a serious complication.

4 Lower bounds

In this section, minimax lower bounds are given. Location estimation is addressed in Section 4.1 and variance estimation is discussed in Section 4.2.

4.1 Location estimation

Since adaptation to σ^2 when estimating the location parameter is possible, we will prove the lower bounds as if $\sigma^2 = 1$ were known. Recall that $P_{\theta,\gamma}$ and $E_{\theta,\gamma}$ are used in place of $P_{\theta,\gamma,1}$ and $E_{\theta,\gamma,1}$ respectively. The Fourier transform of an integrable function f is $\hat{f}(t) = \int e^{-itx} f(x)\,dx$ and the Fourier transform of a finite Borel measure π is $\hat{\pi}(t) = \int e^{-itx}\pi(dx)$.

A minimax lower bound can be established for $1 \le k \le \frac{n}{2} - \sqrt{n}$ via a Fourier-based approach, showcasing an interesting parallel with the upper bound involving a Fourier-based estimator proposed in Section 2.

Theorem 4.1. *Suppose $1 \le k \le \frac{n}{2} - \sqrt{n}$. There exist some universal constants $C, c > 0$ such that*

$$\inf_{\hat{\theta}} \sup_{\substack{\theta \in \mathbb{R} \\ ||\gamma||_0 \le k}} P_{\theta,\gamma}\left\{|\hat{\theta} - \theta| > C \cdot \frac{k}{n}\log^{-1/2}\left(1 + \frac{k^2(n-2k)^2}{n^3}\right)\right\} \ge c.$$

Note Theorem 4.1 implies consistent estimation is impossible when $n - 2k \lesssim \sqrt{n}$ as at least constant order error is unavoidable. Furthermore, consider

$$\frac{k}{n\sqrt{\log\left(1 + \frac{k^2(n-2k)^2}{n^3}\right)}} \asymp \begin{cases} \frac{1}{\sqrt{n}} & \text{if } k \le \sqrt{n}, \\ \frac{k}{n\sqrt{\log\left(1 + \frac{k^2(n-2k)^2}{n^3}\right)}} & \text{if } \sqrt{n} < k < \frac{n}{2} - \sqrt{n}, \end{cases}$$

from the inequality $u/2 \le \log(1 + u) \le u$ for $u \in (0,1)$, and so the lower bound does indeed match the upper bound. The minimax lower bound for $k > \frac{n}{2} - \sqrt{n}$ will be presented in Section 5.1.2.

The proof of Theorem 4.1 proceeds by first linking the model (1) to its mixture formulation (7) via a standard concentration argument. The lower bound argument then involves considering the testing problem

$$H_0 : X_1, ..., X_n \overset{iid}{\sim} (1 - \varepsilon)N(-\theta, 1) + \varepsilon(Q_0 * N(0, 1)),$$
$$H_1 : X_1, ..., X_n \overset{iid}{\sim} (1 - \varepsilon)N(\theta, 1) + \varepsilon(Q_1 * N(0, 1)).$$

where contamination distributions Q_0 and Q_1 as well as the location parameter $\theta \in \mathbb{R}$ are all to be selected. Here, $\varepsilon = \frac{k}{n}$. As usual in minimax lower bound arguments, the goal is to construct Q_0 and Q_1 such that θ can be taken as large as possible while ensuring H_0 and H_1 cannot be distinguished. It turns out a Fourier-based approach [7, 10, 17] yields a rate-optimal construction. Letting f_0 and f_1 denote the marginal densities of H_0 and H_1 respectively, the parameters Q_0, Q_1, and θ are selected such that the Fourier transforms of f_0 and f_1 agree on as wide an interval $[-\tau, \tau]$ as possible.

As argued in the literature [7, 17, 10], the χ^2-divergence admits a bound in terms of the Fourier transforms of the marginal densities,

$$\chi^2(f_1 \,||\, f_0) \leq \frac{\sqrt{2\pi}}{1-\varepsilon} \sum_{k=0}^{\infty} \frac{1}{2^k k!} \int_{-\infty}^{\infty} \left| \frac{d^k}{dt^k} \left(\hat{f}_1(t) - \hat{f}_0(t) \right) e^{-it\theta} \right|^2 dt.$$

If Q_0, Q_1, and θ are chosen such that the Fourier transforms match $\hat{f}_1(t) = \hat{f}_0(t)$ on $[-\tau, \tau]$, then there is hope for the above integral to be small. It turns out the optimal choice of τ in Theorem 2.2 is precisely the correct choice in the lower bound as well, and essentially the optimal choice for the location parameter is $|\theta| \asymp \frac{\varepsilon}{\tau}$.

Let us elaborate by making some more technical remarks. The condition $\hat{f}_0(t) = \hat{f}_1(t)$ implies one should pick Q_0 and Q_1 such that

$$\hat{Q}_1(t) - \hat{Q}_0(t) = \frac{1-\varepsilon}{\varepsilon} \cdot 2i \sin(t\theta).$$

Since $\varepsilon < \frac{1}{2}$ implies $\frac{1-\varepsilon}{\varepsilon} > 1$, it is clear that this cannot hold for all $t \in \mathbb{R}$ since $|\hat{Q}_1(t) - \hat{Q}_0(t)| \leq 2$ because $|\hat{Q}_1(t)| \vee |\hat{Q}_0(t)| \leq 1$ by virtue of Q_0 and Q_1 being probability measures. An idea is to seek Q_0 and Q_1 such that

$$\hat{Q}_1(t) - \hat{Q}_0(t) = \begin{cases} 0 & \text{if } t < -2\tau, \\ \frac{1-\varepsilon}{\varepsilon} \cdot 2i \sin(\tau\theta) \cdot \frac{-2\tau - t}{\tau} & \text{if } -2\tau \leq t \leq -\tau, \\ \frac{1-\varepsilon}{\varepsilon} \cdot 2i \sin(t\theta) & \text{if } -\tau < t < \tau, \\ \frac{1-\varepsilon}{\varepsilon} \cdot 2i \sin(\tau\theta) \cdot \frac{2\tau - t}{\tau} & \text{if } \tau \leq t \leq 2\tau, \\ 0 & \text{if } t > 2\tau, \end{cases} \quad (23)$$

where τ is a tuning parameter to be set and will depend on the choice of θ. The linear functions for $\tau \leq |t| \leq 2\tau$ simply taper to zero. When $|\theta| \asymp \frac{\varepsilon}{\tau}$, the existence of such Q_0 and Q_1 is given by Proposition B.1.

Though appealing in its simplicity, it seems to us this avenue is only capable of delivering a sharp minimax lower bound for, say, the regime $\varepsilon \leq \frac{1}{4}$. The behavior for ε close to $\frac{1}{2}$ is missed; indeed, it is not clear from this idea why the term $1 - 2\varepsilon \left(= \frac{n-2k}{n} \right)$ should even appear in the rate. Roughly speaking, choosing $|\theta| \asymp \frac{\varepsilon}{\tau}$ yields the following bound on the χ^2-divergence

$$\chi^2(f_0 \,||\, f_1) \lesssim \varepsilon^2 e^{-C\tau^2},$$

where $C > 0$ is a universal constant. To have $\chi^2(f_0 \,||\, f_1) \lesssim \frac{1}{n}$ would force the choice $\tau \asymp \sqrt{\log(en\varepsilon^2)} \asymp \sqrt{\log\left(\frac{ek^2}{n}\right)}$, thus missing the $n - 2k$ scaling.

The main issue is that the approximation of $\frac{1-\varepsilon}{\varepsilon} \cdot 2i \sin(t\theta)$ by 0 for $|t| > 2\tau$ is not good enough. Instead, consider seeking for $|t| > 2\tau$,

$$\hat{Q}_1(t) - \hat{Q}_0(t) = 2\varepsilon \cdot 2i \sin(t\theta). \quad (24)$$

Observe that now the approximation error is $\left| \frac{1-\varepsilon}{\varepsilon} (2i \sin(t\theta)) - 2\varepsilon (2i \sin(t\theta)) \right| = (1-2\varepsilon)\frac{(1+\varepsilon)}{\varepsilon} |2i \sin(t\theta)|$, which scales with $1 - 2\varepsilon$. Of course, a major question is whether it is even possible to find two probability measures Q_0 and Q_1 satisfying (24). It turns out via a simple reparameterization that this question can be reduced to asking the same question in the context of the previous idea, namely whether there exists a pair $\tilde{Q}_0, \tilde{Q}_1$ satisfying (23). Consider the parameterization

$$Q_0 = 2\varepsilon \delta_\theta + (1 - 2\varepsilon)\tilde{Q}_0,$$
$$Q_1 = 2\varepsilon \delta_{-\theta} + (1 - 2\varepsilon)\tilde{Q}_1.$$

Then $\hat{Q}_1(t) - \hat{Q}_0(t) = 2\varepsilon \cdot 2i \sin(t\theta) + (1 - 2\varepsilon)\left(\hat{\tilde{Q}}_1(t) - \hat{\tilde{Q}}_0(t) \right)$. Then both the condition (24)

for $|t| \geq 2\tau$ and the condition $\hat{Q}_1(t) - \hat{Q}_0(t) = \frac{1-\varepsilon}{\varepsilon} \cdot 2i\sin(t\theta)$ for $|t| \leq \tau$ is satisfied if $\tilde{Q}_0, \tilde{Q}_1$ satisfy (23) with contamination proportion $\frac{\varepsilon}{1+2\varepsilon}$ in place of ε. The choice of parameterization admits a natural explanation. In the case $\varepsilon = \frac{1}{2}$, we have $Q_0 = \delta_\theta$ and $Q_1 = \delta_{-\theta}$. This is natural as the location parameter is not identifiable and $\chi^2(f_1 \,||\, f_0) = 0$. This extreme case suggests the above parameterization, which interpolates between the construction (23) (useful for $\varepsilon < \frac{1}{4}$) and the extremal case $\varepsilon = \frac{1}{2}$.

The technical parts of the proof consist entirely in constructing suitable $\tilde{Q}_0$ and $\tilde{Q}_1$ and bounding the induced χ^2-divergence using arguments inspired by [17]. The correct choices of τ and θ emerge as consequences. Roughly speaking, choosing $|\theta| \asymp \frac{\varepsilon}{\tau}$ yields the following bound on the χ^2-divergence

$$\chi^2(f_0 \,||\, f_1) \lesssim (1 - 2\varepsilon)^2 \varepsilon^2 e^{-C\tau^2},$$

where $C > 0$ is a universal constant. Then the correct choice $\tau \asymp \sqrt{\log(1 + n\varepsilon^2(1-2\varepsilon)^2)} \asymp \sqrt{\log\left(1 + \frac{k^2(n-2k)^2}{n^3}\right)}$ is available to yield $\chi^2(f_0 \,||\, f_1) \lesssim \frac{1}{n}$. This rough sketch captures the essential spirit of the proof of Theorem 4.1.

4.2 Variance estimation

For variance estimation, a minimax lower bound is given by Part (ii) of Proposition 7 in [17], matching the upper bound of Section 3.2. We give a statement of their lower bound in the present context for completeness.

Theorem 4.2 (Part (ii) of Proposition 7 in [17]). *There exist some universal constants $C, c > 0$ such that*

$$\inf_{\hat{\theta}} \sup_{\substack{\theta \in \mathbb{R} \\ ||\gamma||_0 \leq k \\ \sigma > 0}} P_{\theta, \gamma, \sigma} \left\{ \frac{|\hat{\sigma}^2 - \sigma^2|}{\sigma^2} > \frac{Ck}{n\log\left(1 + \frac{k}{\sqrt{n}}\right)} \right\} \geq c.$$

5 Discussion

In this section, some remarks are made discussing finer points of interest.

5.1 Inconsistent, yet rate-optimal estimation for $n - 2k \lesssim \sqrt{n}$

In the regime $n - 2k \lesssim \sqrt{n}$, consistent estimation of θ is impossible as implied by the lower bound of Section 4.1. From the perspective of large-scale inference which motivates the two-groups model (1), it seems uninteresting to investigate this regime further. However, it turns out that the minimax rate in this regime still exhibits a subtle logarithmic dependence on $n - 2k$, which can be derived from recent understanding of a kernel mode estimator [38].

5.1.1 Upper bound

Motivated by the recent results in [38], we consider a kernel mode estimator with a box kernel. For a bandwidth $h > 0$, define $\hat{G}_h(t) := \frac{1}{2nh} \sum_{j=1}^n \mathbb{1}_{\{|t - X_j| \leq h\}}$ and set

$$\hat{\theta} := \operatorname*{argmax}_{t \in \mathbb{R}} \hat{G}_h(t). \tag{25}$$

To simplify presentation, the variance will be assumed known and equal to one in this section; adaptation to σ^2, as well as k, has been discussed in [38].

A mode-type estimator is natural for the purpose of null distribution estimation. In the setting of two-groups model (1), the majority of the coordinates share the same mean value θ. Efron [23, 25] essentially estimates a marginal density of the z-scores and then takes the center of the peak as an estimator of θ. The kernel mode estimator in the form of (25) has a long history. It was first proposed

by Parzen [50] in his original paper of kernel density estimation. Consistency, rates of convergence and asymptotic distribution were investigated by Chernoff [14] and Eddy [22].

The recent paper [38] studies the statistical property of the kernel mode estimator in the context of robust estimation. Unlike the classical literature [50, 14, 22], it turns out the optimal choice of the bandwidth is a widening, rather than shrinking, one.

Theorem 5.1 (Theorem 3.1 [38]). *Suppose* $1 \leq k < \frac{n}{2}$. *There exist universal constants* $C_1, C_2 > 0$ *such that the following holds. For any* $\delta \in (0,1)$, *there exists* L_δ *depending only on* δ *such that if* n *is sufficiently large depending only on* δ *and* $h = C_1 \sqrt{1 \vee \log (L_\delta n/(n-2k)^2)}$, *then*

$$\sup_{\substack{\theta \in \mathbb{R} \\ ||\gamma||_0 \leq k}} P_{\theta,\gamma} \left\{ |\hat{\theta} - \theta| > C_2 h \right\} \leq \delta$$

where $\hat{\theta}$ *is given by* (25).

Applying Theorem 5.1 to the regime $\frac{n}{2} - C\sqrt{n} < k < \frac{n}{2}$, the minimax rate (11) is achieved in the inconsistency regime, thus complementing the result of Theorem 2.2 for the Fourier-based estimator.

5.1.2 Lower bound

In Section 5.1.1, a kernel mode estimator is used to achieve the minimax rate in the inconsistency regime. The following matching minimax lower bound can be established, which is stated for $\sigma^2 = 1$ without loss of generality.

Theorem 5.2. *Suppose* $\frac{n}{2} - \sqrt{n} < k < \frac{n}{2}$. *There exist some universal constants* $C, c > 0$ *such that*

$$\inf_{\hat{\theta}} \sup_{\substack{\theta \in \mathbb{R} \\ ||\gamma||_0 \leq k}} P_{\theta,\gamma} \left\{ |\hat{\theta} - \theta| > C \sqrt{\log \left(1 + \frac{n}{(n-2k)^2} \right)} \right\} \geq c.$$

Together with the result of Theorem 4.1, we have established the lower bound of the minimax rate (11) for all $1 \leq k < \frac{n}{2}$.

The construction in the proof of Theorem 5.2 is, at its core, similar to that in [38]. Since the context of [38] is quite different, we still present a self-contained proof in the paper. Instead of constructing contamination distributions and selecting a location parameter such that the characteristic functions of the marginals match on a large interval containing zero, a more direct argument is available. Strictly speaking, the concentration argument used in Section 4.1 to link the frequentist model (1) to the Bayes model (7) can no longer be used. The proof of Theorem 5.2 is thus directly constructed for the frequentist model (1).

5.2 Null estimation in total variation

In some situations, estimation of the null distribution with respect to some information-theoretic distance or divergence is of interest. Our results regarding parameter estimation turn out to directly yield minimax rates for estimating the null distribution in the total variation distance. The connection between the total variation of two univariate Gaussian distributions and their parameters is given by the following result of Devroye, Mehrabian, and Reddad [19].

Lemma 5.3 (Theorem 1.3 in [19]). *If* $\mu_1, \mu_2 \in \mathbb{R}$ *and* $\sigma_1, \sigma_2 > 0$, *then*

$$d_{\mathrm{TV}} \left(N(\mu_1, \sigma_1^2), N(\mu_2, \sigma_2^2) \right) \asymp 1 \wedge \left(\frac{|\sigma_1^2 - \sigma_2^2|}{\sigma_1^2 \vee \sigma_2^2} \vee \frac{|\mu_1 - \mu_2|}{\sigma_1 \vee \sigma_2} \right).$$

This result immediately yields the minimax rate for estimation of the null distribution in total variation, namely

$$\epsilon^*_{\mathrm{TV}}(k,n) \asymp 1 \wedge \frac{k}{n \sqrt{\log \left(1 + \frac{k^2(n-2k)^2}{n^3} \right)}} \asymp \begin{cases} \frac{k}{n} \log^{-\frac{1}{2}} \left(1 + \frac{k^2(n-2k)^2}{n^3} \right) & \text{if } 1 \leq k \leq \frac{n}{2} - \sqrt{n}, \\ 1 & \text{if } \frac{n}{2} - \sqrt{n} < k < \frac{n}{2}. \end{cases}$$

Notably, the location estimation rate dominates and thus determines the total variation rate. A formal statement of the minimax rate follows, which we give without proof.

Theorem 5.4. *Suppose $1 \leq k < \frac{n}{2}$, and consider the estimators $\hat{\theta}$ and $\hat{\sigma}^2$ from Theorems 3.2 and 3.4, respectively. For any $\delta \in (0, 1)$, there exists some constant $C > 0$ depending only on δ such that*

$$\sup_{\substack{\theta \in \mathbb{R} \\ ||\gamma||_0 \leq k \\ \sigma > 0}} P_{\theta, \gamma, \sigma} \left\{ \mathrm{d_{TV}} \left(N(\hat{\theta}, \hat{\sigma}^2), N(\theta, \sigma^2) \right) > C \left(\frac{k}{n\sqrt{\log \left(1 + \frac{k^2(n-2k)^2}{n^3} \right)}} \wedge 1 \right) \right\} \leq \delta.$$

Furthermore, there exist some universal constants $C', c > 0$ such that

$$\inf_{\hat{\theta}, \hat{\sigma}} \sup_{\substack{\theta \in \mathbb{R} \\ ||\gamma||_0 \leq k \\ \sigma > 0}} P_{\theta, \gamma, \sigma} \left\{ \mathrm{d_{TV}} \left(N(\hat{\theta}, \hat{\sigma}^2), N(\theta, \sigma^2) \right) \geq C' \left(\frac{k}{n\sqrt{\log \left(1 + \frac{k^2(n-2k)^2}{n^3} \right)}} \wedge 1 \right) \right\} \geq c.$$

Since for $k \geq \frac{n}{2} - \sqrt{n}$ the lower bound is $\epsilon_{\mathrm{TV}}^*(k, n) \gtrsim 1$, it follows from the boundedness of total variation that $N(\hat{\theta}, \hat{\sigma}^2)$ achieves the minimax rate even though no guarantees are available for $\hat{\theta}$ when $n - 2k \lesssim \sqrt{n}$. We always have $\mathrm{d_{TV}} \left(N(\hat{\theta}, \hat{\sigma}^2), N(\theta, \sigma^2) \right) \lesssim 1$. This marks a significant difference between parameter estimation (in which minimax estimation for $k \geq \frac{n}{2} - \sqrt{n}$ is nontrivial) and estimation in total variation.

5.3 Adaptation to k

Thus far it has been assumed that the signal/contamination level k in (1) is known to the statistician. In practice, this is usually not the case and so adaptation to unknown k is an important problem. As frequently seen in the literature, a standard application of Lepski's method [42, 43, 44] gives an adaptive estimator of the location parameter θ. Fix $\delta, \eta \in (0, 1)$. We will consider adaptation to k for $1 \leq k \leq \frac{n}{2} - C_\delta \sqrt{n}$ where $C_\delta > 0$ is the constant depending on δ from Theorem 3.4. Let $\tilde{\sigma}^2$ denote the pilot variance estimator in Proposition 3.1 at confidence level η. By Proposition 3.1 there exists $L_\eta \geq 1$ such that $\tilde{\sigma}^2 L_\eta^{-1} \leq \sigma^2 \leq \tilde{\sigma}^2 L_\eta$ with probability at least $1 - \eta$ uniformly over the parameter space. To apply Lepski's method, define

$$\mathcal{K} := \left\{ 1, 2, 3, ..., \left\lfloor \frac{n}{2} - C_\delta \sqrt{n} \right\rfloor \right\}. \tag{26}$$

Let $\hat{\theta}_k$ denote the estimator of Theorem 3.4 at signal level k and confidence level δ. Define

$$\epsilon(k, n) = \frac{k}{n\sqrt{\log \left(1 + \frac{k^2(n-2k)^2}{n^3} \right)}}. \tag{27}$$

For $k \in \mathcal{K}$, define the intervals

$$J_k = \left[\hat{\theta}_k - C'_{\delta, \eta} \sqrt{L_\eta} \tilde{\sigma} \epsilon(k, n), \hat{\theta}_k + C'_{\delta, \eta} \sqrt{L_\eta} \tilde{\sigma} \epsilon(k, n) \right],$$

where $C'_{\delta, \eta}$ is chosen large enough depending on δ and η. The adaptive estimator $\hat{\theta}$ is defined to be any element of the set

$$\bigcap_{\substack{k \in \mathcal{K} \\ k \geq k'}} J_k, \tag{28}$$

where $k' \in \mathcal{K}$ is the smallest value such that the set is nonempty. If no such k' exists, set $\hat{\theta} = 0$. Note that computing $\hat{\theta}$ does not require any knowledge of the true signal/contamination level nor the true variance.

Theorem 5.5. *Suppose $\delta, \eta \in (0, 1)$. There exists $C_\delta, C_{\delta, \eta} > 0$ sufficiently large depending only on δ and δ, η respectively such that the following holds. If n is sufficiently large depending only on δ, η and*

18

$1 \le k^* \le \frac{n}{2} - C_\delta \sqrt{n}$, then

$$\sup_{\substack{\theta \in \mathbb{R} \\ ||\gamma||_0 \le k^* \\ \sigma > 0}} P_{\theta, \gamma, \sigma} \left\{ \frac{|\hat{\theta} - \theta|}{\sigma} > C_{\delta, \eta} \epsilon(k^*, n) \right\} \le \eta + \delta,$$

where $\epsilon(k, n)$ is given by (27).

Thus, the adaptive estimator $\hat{\theta}$ achieves the minimax rate. It is not so surprising that adaptation to k is possible since it has been established in a one-sided contamination model [9]. Of course, adaptation is trivial in Huber's contamination model since the sample median is already rate-optimal.

The construction of an adaptive variance estimator is similar. Fix $\delta, \eta \in (0, 1)$. Recall that $\tilde{\sigma}^2$ is a pilot variance estimator satisfying $\tilde{\sigma}^2 L_\eta^{-1} \le \sigma^2 \le \tilde{\sigma}^2 L_\eta$ with probability at least $1 - \eta$ uniformly over the parameter space. Let $\hat{\sigma}_k^2$ denote the estimator from Section 3.2 at confidence level δ and signal level k. Define

$$\epsilon_{\text{var}}(k, n) = \frac{k}{n \log \left(1 + \frac{k}{\sqrt{n}} \right)}. \tag{29}$$

For $1 \le k < \frac{n}{2}$, define the intervals

$$J_k = \left[\hat{\sigma}_k^2 - \tilde{\sigma}^2 C'_{\delta, \eta} L_\eta \epsilon_{\text{var}}(k, n), \hat{\sigma}_k^2 + \tilde{\sigma}^2 C'_{\delta, \eta} L_\eta \epsilon_{\text{var}}(k, n) \right],$$

where $C'_{\delta, \eta}$ is a sufficiently large constant depending only on η and δ. The adaptive estimator $\hat{\sigma}^2$ is defined to be any element of the set

$$\bigcap_{k' \le k} J_k, \tag{30}$$

where $1 \le k' < \frac{n}{2}$ is the smallest value such that the set is nonempty. If no such k' exists, set $\hat{\sigma}^2 = 1$. Again, note that computing $\hat{\sigma}^2$ does not require any knowledge of the true signal/contamination level. The following guarantee is available.

Theorem 5.6. *Suppose $1 \le k^* < \frac{n}{2}$. If $\delta, \eta \in (0, 1)$, and n is sufficiently large depending only on η and δ, then there exists $C_{\delta, \eta} > 0$ depending only on η and δ such that*

$$\sup_{\substack{\theta \in \mathbb{R} \\ ||\gamma||_0 \le k^* \\ \sigma > 0}} P_{\theta, \gamma, \sigma} \left\{ \frac{|\hat{\sigma}^2 - \sigma^2|}{\sigma^2} > C_{\delta, \eta} \epsilon_{var}(k^*, n) \right\} \le \eta + \delta.$$

With an adaptive variance estimator in hand which can achieve the minimax rate of location estimation, an adaptive estimator of the null distribution in total variation distance can be immediately constructed, yielding the following result which we state without proof.

Theorem 5.7. *For $\delta \in (0, 1)$, there exists $\tilde{C} > 0$ depending only on δ such that the following holds. If n is sufficiently large depending only on δ, then*

$$\sup_{\substack{\theta \in \mathbb{R} \\ ||\gamma||_0 \le k \\ \sigma > 0}} P_{\theta, \gamma, \sigma} \left\{ d_{\text{TV}} \left(N(\hat{\theta}, \hat{\sigma}^2), N(\theta, \sigma^2) \right) > \tilde{C} \left(\frac{k}{n \sqrt{\log \left(1 + \frac{k^2 (n - 2k)^2}{n^3} \right)}} \wedge 1 \right) \right\} \le \delta$$

with appropriate hyperparameters for the adaptive estimators $\hat{\theta}$ and $\hat{\sigma}^2$.

Thus, the minimax rate can be achieved without knowledge of k.

5.4 General noise distributions

The model (1) stipulates the noise is Gaussian. However, it is also interesting to consider the generic version of (1) where $Z_1, ..., Z_n \overset{iid}{\sim} F$ for some known symmetric distribution F (e.g. Cauchy, Laplace, etc.). For discussion, take $\sigma^2 = 1$ to be known.

Since the parametric rate can always be achieved by the sample median for $k \lesssim \sqrt{n}$, for sake of discussion consider the regime $\sqrt{n} \le k \le \frac{n}{2} - C\sqrt{n}$. Our methodology requires only a slight

modification to accommodate noise variables drawn from F. Note in (14) the term $e^{\frac{\omega^2}{2}}$ is the reciprocal of the characteristic function of $N(0,1)$ evaluated at ω. Therefore, a natural idea is to consider the estimator

$$\hat{\theta}_F = \operatorname*{argmin}_{\mu \in \mathbb{R}} \sup_{|\omega| \le \tau} \inf_{\substack{\zeta \in \mathbb{C} \\ |\zeta| \le 1}} \left| \frac{1}{n} \sum_{j=1}^{n} \frac{e^{i\omega(X_j - \mu)}}{\psi_F(\omega)} - \frac{n-k}{n} - \frac{k}{n}\zeta \right|,$$

where $\psi_F(\omega) = E(e^{i\omega Z_1})$ is the characteristic function of F.

The arguments in the Gaussian case can be straightforwardly modified to show that with the choice τ such that $\frac{1}{\psi_F(\tau)} \asymp \frac{k(n-2k)}{n^{3/2}}$, we have $|\hat{\theta}_F - \theta| \lesssim \frac{k}{n}\tau^{-1}$ with high probability. For example, if F is the Laplace distribution with unit scale, then $\psi_F(\omega) = \frac{1}{1+\omega^2}$, and so the choice $\tau^2 \asymp \frac{k(n-2k)}{n^{3/2}}$ yields $|\hat{\theta}_F - \theta|^2 \lesssim \frac{1}{\sqrt{n}} \cdot \frac{k}{(n-2k)}$. Notably, this Laplace rate is faster than the Gaussian rate (11) by a polynomial factor! The improvement in the rate can be seen as the result of the Laplace distribution having a characteristic function which decays polynomially in ω rather than exponentially like that of the standard Gaussian distribution. This phenomenon has been noted before in the deconvolution literature [46, 26, 11]. An interesting problem is to prove a matching minimax lower bound for estimating θ with a generic noise distribution F.

6 Proofs

This section contains the analyses of the Fourier-based location estimator of Section 2 and the variance estimator of Section 3.2.

6.1 Fourier-based estimator: known variance

Proof of Theorem 2.2. Fix $\delta \in (0,1)$. It follows from an argument using the bounded differences inequality (e.g. Theorem D.13) that for a sufficiently large $L > 0$ depending only on δ, the event $\mathcal{E} := \left\{ \sup_{\omega \in \mathbb{R}} \left| \frac{1}{n} \sum_{j=1}^{n} \left(e^{i\omega(X_j - \theta)} - E_{\theta,\gamma}\left(e^{i\omega(X_j - \theta)} \right) \right) \right| \le \frac{L}{\sqrt{n}} \right\}$ has $P_{\theta,\gamma}$-probability of at least $1 - \delta$ uniformly over (θ, γ). Examining (14), by definition of $\hat{\theta}$ we have on the event $\mathcal{E}$,

$$\sup_{|\omega| \le \tau} \inf_{\substack{\zeta \in \mathbb{C}, \\ |\zeta| \le 1}} \left| \frac{1}{n} \sum_{j=1}^{n} e^{i\omega(X_j - \hat{\theta}) + \frac{\omega^2}{2}} - \frac{n-k}{n} - \frac{k}{n}\zeta \right|$$

$$\le \sup_{|\omega| \le \tau} \inf_{\substack{\zeta \in \mathbb{C}, \\ |\zeta| \le 1}} \left| \frac{1}{n} \sum_{j=1}^{n} e^{i\omega(X_j - \theta) + \frac{\omega^2}{2}} - \frac{n-k}{n} - \frac{k}{n}\zeta \right|$$

$$\le \sup_{|\omega| \le \tau} \left| \frac{1}{n} \sum_{j=1}^{n} e^{i\omega(X_j - \theta) + \frac{\omega^2}{2}} - \frac{n-k}{n} - \frac{k}{n} \cdot \frac{1}{k} \sum_{j \in \mathcal{O}} e^{i\omega\gamma_j} \right|$$

$$= \sup_{|\omega| \le \tau} \left| \frac{1}{n} \sum_{j=1}^{n} \left(e^{i\omega(X_j - \theta) + \frac{\omega^2}{2}} - E_{\theta,\gamma}\left(e^{i\omega(X_j - \theta) + \frac{\omega^2}{2}} \right) \right) \right|$$

$$\le \frac{Le^{\frac{\tau^2}{2}}}{\sqrt{n}}. \tag{31}$$

Now, consider that on $\mathcal{E}$ we also have

$$\sup_{|\omega|\leq\tau}\left|\frac{1}{n}\sum_{j=1}^{n}\left(e^{i\omega(X_j-\hat\theta)+\frac{\omega^2}{2}}-e^{i\omega(\theta+\gamma_j-\hat\theta)}\right)\right|=\sup_{|\omega|\leq\tau}\left|e^{-i\omega\hat\theta+\frac{\omega^2}{2}}\right|\left|\frac{1}{n}\sum_{j=1}^{n}\left(e^{i\omega X_j}-e^{i\omega(\theta+\gamma_j)-\frac{\omega^2}{2}}\right)\right|$$

$$\leq e^{\frac{\tau^2}{2}}\cdot\sup_{\omega\in\mathbb{R}}\left|\frac{1}{n}\sum_{j=1}^{n}\left(e^{i\omega X_j}-E_{\theta,\gamma}\left(e^{i\omega X_j}\right)\right)\right|$$

$$=e^{\frac{\tau^2}{2}}\cdot\sup_{\omega\in\mathbb{R}}\left|\frac{1}{n}\sum_{j=1}^{n}\left(e^{i\omega(X_j-\theta)}-E_{\theta,\gamma}\left(e^{i\omega(X_j-\theta)}\right)\right)\right|$$

$$\leq\frac{Le^{\frac{\tau^2}{2}}}{\sqrt{n}}.$$

Therefore, by reverse triangle inequality, it follows that on $\mathcal{E}$

$$\sup_{|\omega|\leq\tau}\inf_{\substack{\zeta\in\mathbb{C},\\|\zeta|\leq 1}}\left|\frac{1}{n}\sum_{j=1}^{n}e^{i\omega(X_j-\hat\theta)+\frac{\omega^2}{2}}-\frac{n-k}{n}-\frac{k}{n}\zeta\right|$$

$$\geq\sup_{|\omega|\leq\tau}\inf_{\substack{\zeta\in\mathbb{C},\\|\zeta|\leq 1}}\left|\frac{1}{n}\sum_{j=1}^{n}e^{i\omega(\theta+\gamma_j-\hat\theta)}-\frac{n-k}{n}-\frac{k}{n}\zeta\right|-\frac{Le^{\frac{\tau^2}{2}}}{\sqrt{n}}$$

$$=\sup_{|\omega|\leq\tau}\inf_{\substack{\zeta\in\mathbb{C},\\|\zeta|\leq 1}}\left|\frac{n-k}{n}\left(e^{i\omega(\theta-\hat\theta)}-1\right)+\frac{k}{n}\left(\frac{1}{k}\sum_{j\in\mathcal{O}'}e^{i\omega(\theta+\gamma_j-\hat\theta)}-\zeta\right)\right|-\frac{Le^{\frac{\tau^2}{2}}}{\sqrt{n}}$$

$$\geq\sup_{|\omega|\leq\tau}\inf_{\substack{z\in\mathbb{C},\\|z|\leq 2}}\left|\frac{n-k}{n}\left(e^{i\omega(\theta-\hat\theta)}-1\right)+\frac{k}{n}z\right|-\frac{Le^{\frac{\tau^2}{2}}}{\sqrt{n}}. \tag{32}$$

Here, $\mathcal{O}'$ is defined to be $\mathcal{O}$ with any arbitrary $k-|\mathcal{O}|$ indices added from $\mathcal{I}$. This is simply to ensure $|\mathcal{O}'|=k$. Let us now examine the first term on the right hand side of (32). We break up the analysis into two cases to separately obtain the parametric part $\frac{1}{\sqrt{n}}$ and the nonparametric part $\frac{k}{n\tau}$ in our target bound. Let $C_1>0$ be sufficiently large depending only on δ.

Case 1: Suppose $k\leq C_1\sqrt{n}$. The argument is fairly simple to obtain the parametric rate. From (32) and (31) it follows

$$\sup_{|\omega|\leq\tau}\left|e^{i\omega(\theta-\hat\theta)}-1\right|\leq\left(\frac{n}{n-k}\right)\left(\frac{2Le^{\frac{\tau^2}{2}}}{\sqrt{n}}+\frac{2k}{n}\right)\leq\frac{1}{\sqrt{n}}\left(\frac{1}{1-\frac{C_1}{\sqrt{n}}}\right)\left(2Le^{\frac{\tau^2}{2}}+2C_1\right)\leq\frac{\tilde{C}}{\sqrt{n}}$$

where $\tilde{C}>0$ depends only on δ. Here, we have used n is sufficiently large (depending only on δ). We have also used $k\leq C_1\sqrt{n}$ implies $\log\left(1+\frac{k^2(n-2k)^2}{n^3}\right)\leq\log\left(1+C_1^2\right)$, and so $\tau=1$ by selecting c sufficiently small depending only on δ. Let $\tilde{C}'=\frac{12}{5}\tilde{C}$. We claim $|\hat\theta-\theta|\leq\frac{\tilde{C}'}{\sqrt{n}}$. To prove the claim, suppose not, that is, suppose $|\hat\theta-\theta|>\frac{\tilde{C}'}{\sqrt{n}}$. Consider the choice $\omega^*=\frac{\tilde{C}'}{\sqrt{n}(\theta-\hat\theta)}$ and observe $|\omega^*|\leq 1=\tau$. Therefore, from the inequality $\frac{5}{6}x\leq\sin x\leq|e^{ix}-1|$ for $x\in(0,1)$, we have

$$\frac{2\tilde{C}}{\sqrt{n}}=\frac{5}{6}\cdot\frac{\tilde{C}'}{\sqrt{n}}=\frac{5}{6}\omega^*(\theta-\hat\theta)\leq\left|e^{i\omega^*(\theta-\hat\theta)}-1\right|\leq\sup_{|\omega|\leq\tau}\left|e^{i\omega(\theta-\hat\theta)}-1\right|\leq\frac{\tilde{C}}{\sqrt{n}}$$

which is a contradiction. Hence, we have the claim. To summarize, we have shown $|\hat\theta-\theta|\leq\frac{\tilde{C}'}{\sqrt{n}}$ on the event $\mathcal{E}$ when $k\leq C_1\sqrt{n}$, as desired.

Case 2: Suppose $k>C_1\sqrt{n}$. For any fixed ω with $|\omega|\leq\tau$, it is straightforward to solve the optimization problem

$$\inf_{\substack{z\in\mathbb{C},\\|z|\leq 2}}\left|\frac{n-k}{n}\left(e^{i\omega(\theta-\hat\theta)}-1\right)+\frac{k}{n}z\right|$$

21

directly and obtain the solution

$$
z^* = \begin{cases}
-\frac{n-k}{k}\left(e^{i\omega(\theta-\hat{\theta})}-1\right) & \text{if } \left|e^{i\omega(\theta-\hat{\theta})}-1\right| \leq \frac{2k}{n-k}, \\
-\frac{2}{\left|e^{i\omega(\theta-\hat{\theta})}-1\right|}\cdot\left(e^{i\omega(\theta-\hat{\theta})}-1\right) & \text{otherwise.}
\end{cases}
$$

Plugging in z^*, we see the value of the optimization problem is zero when $\left|e^{i\omega(\theta-\hat{\theta})}-1\right| \leq \frac{2k}{n-k}$. Therefore,

$$
\sup_{|\omega|\leq\tau}\inf_{\substack{z\in\mathbb{C},\\|z|\leq 2}}\left|\frac{n-k}{n}\left(e^{i\omega(\theta-\hat{\theta})}-1\right)+\frac{k}{n}z\right| = \sup_{|\omega|\leq\tau}\left|\frac{n-k}{n}\left(e^{i\omega(\theta-\hat{\theta})}-1\right)+\frac{k}{n}z^*\right|
$$

$$
= \sup_{\substack{|\omega|\leq\tau,\\\left|e^{i\omega(\theta-\hat{\theta})}-1\right|>\frac{2k}{n-k}}}\left|\frac{n-k}{n}\left(e^{i\omega(\theta-\hat{\theta})}-1\right)+\frac{k}{n}z^*\right|
$$

$$
= \sup_{\substack{|\omega|\leq\tau,\\\left|e^{i\omega(\theta-\hat{\theta})}-1\right|>\frac{2k}{n-k}}}\left|\frac{n-k}{n}-\frac{2k/n}{|e^{i\omega(\theta-\hat{\theta})}-1|}\right|\left|e^{i\omega(\theta-\hat{\theta})}-1\right|.
$$

To summarize, from (31) and (32) it has been shown that on $\mathcal{E}$ we have

$$
\sup_{\substack{|\omega|\leq\tau,\\\left|e^{i\omega(\theta-\hat{\theta})}-1\right|>\frac{2k}{n-k}}}\left|\frac{n-k}{n}-\frac{2k/n}{|e^{i\omega(\theta-\hat{\theta})}-1|}\right|\left|e^{i\omega(\theta-\hat{\theta})}-1\right| \leq \frac{2Le^{\frac{\tau^2}{2}}}{\sqrt{n}}. \tag{33}
$$

Set $C = \left(\frac{4L}{4-\pi}\right)^2$ and take $c \leq \frac{1}{2}$ (and sufficiently small so that the Case 1 analysis goes through). We claim that (33) implies $|\hat{\theta}-\theta| \leq 8\frac{k}{n}\tau^{-1}$. For sake of contradiction, suppose $|\hat{\theta}-\theta| > \frac{8k}{n}\tau^{-1}$. Now, consider the choice $\omega^* := \frac{\pi\wedge(8k/n)}{\theta-\hat{\theta}}$. Note that $|\omega^*| = \frac{\pi\wedge(8k/n)}{|\hat{\theta}-\theta|} \leq \frac{8}{8}\tau = \tau$. Further consider $|e^{i\omega^*(\theta-\hat{\theta})}-1| = |e^{i(\pi\wedge(8k/n))}-1|$. If $\pi \leq 8k/n$, then $|e^{i\omega^*(\theta-\hat{\theta})}-1| = 2 > \frac{2k}{n-k}$ since $k < \frac{n}{2}$. On the other hand, if $\pi > 8k/n$, then $|e^{i\omega^*(\theta-\hat{\theta})}-1| = |e^{i(8k/n)}-1| \geq \frac{4k}{n} > \frac{2k}{n-k}$. Hence, ω^* lives in the domain of optimization in (33) and so

$$
\frac{2Le^{\frac{\tau^2}{2}}}{\sqrt{n}} \geq \begin{cases}
\left(\frac{n-k}{n}-\frac{1}{2}\right)\left(\frac{4k}{n}\right) & \text{if } \frac{8k}{n} < \pi, \\
2\cdot\frac{n-2k}{n} & \text{if } \frac{8k}{n} \geq \pi
\end{cases}
$$

$$
\geq \begin{cases}
\frac{4-\pi}{2}\cdot\frac{k}{n} & \text{if } \frac{8k}{n} < \pi, \\
2\cdot\frac{n-2k}{n} & \text{if } \frac{8k}{n} \geq \pi.
\end{cases}
$$

$$
\geq \begin{cases}
\frac{4-\pi}{2}\cdot\frac{k}{n}\cdot\frac{n-2k}{n} & \text{if } \frac{8k}{n} < \pi, \\
\frac{4-\pi}{2}\cdot\frac{n-2k}{n}\cdot\frac{k}{n} & \text{if } \frac{8k}{n} \geq \pi.
\end{cases}
$$

$$
= \frac{4-\pi}{2}\cdot\frac{(n-2k)k}{n^2}
$$

where we have used that $\frac{n-2k}{n}$ and $\frac{k}{n}$ are both less than or equal to 1 in the penultimate step. Rearranging, we have

$$
1 \leq \frac{4L}{4-\pi}\cdot\frac{n^{3/2}}{(n-2k)k}e^{\frac{\tau^2}{2}}.
$$

Since $C_1\sqrt{n} \leq k \leq \frac{n}{2}-C\sqrt{n}$, we can take C_1, C large depending only on δ to obtain $\frac{(n-2k)^2k^2}{n^3} \geq C_2$ for a sufficiently large $C_2 > 0$ depending only on δ. Consequently,

$$
\tau^2 = 1 \vee c^2\log\left(1+\frac{(n-2k)^2k^2}{n^3}\right) < \frac{1}{2}\log\left(\frac{(n-2k)^2k^2}{n^3}\right) \leq \log\left(\frac{(n-2k)^2k^2}{n^3}\right)-2\log\left(\frac{4L}{4-\pi}\right).
$$

Therefore, $1 \leq \frac{4L}{4-\pi}\cdot\frac{n^{3/2}}{(n-2k)k}e^{\frac{\tau^2}{2}} < 1$ and so we have a contradiction. Hence, $|\hat{\theta}-\theta| \leq \frac{8k}{n}\tau^{-1}$ on $\mathcal{E}$. The proof is complete. $\qquad\square$

6.2 Variance estimation

Proof of Theorem 3.2. Fix $\delta \in (0,1)$. It follows from an argument using the bounded differences inequality (e.g. Theorem D.13) that for a sufficiently large $L > 0$ depending only on δ, the event $\mathcal{E} := \left\{ \sup_{\omega \in \mathbb{R}} \left| \frac{1}{n} \sum_{j=1}^n \left(e^{i\omega X_j} - E_{\theta,\gamma,\sigma} \left(e^{i\omega X_j} \right) \right) \right| \leq \frac{L}{\sqrt{n}} \right\}$ satisfies $\inf_{\substack{\theta \in \mathbb{R}, \\ ||\gamma||_0 \leq k, \\ \sigma > 0}} P_{\theta,\gamma,\sigma}(\mathcal{E}) \geq 1 - \frac{\delta}{2}$. For $\omega \in \mathbb{R}$, define $N(\omega) := \left| \frac{1}{n} \sum_{j=1}^n e^{i\omega(\theta+\gamma_j) - \frac{\sigma^2 \omega^2}{2}} \right|$. By definition of $\hat{\sigma}^2$, for any $a \leq \omega \leq b$, we have on the event $\mathcal{E}$

$$\hat{\sigma}^2 - \sigma^2 \leq \frac{2}{\omega^2} \left| \log \hat{N}(\omega) - \log N(\omega) \right| - \frac{2 \log N(\omega)}{\omega^2} - \sigma^2$$

$$\leq \frac{2}{\omega^2} \frac{|\hat{N}(\omega) - N(\omega)|}{\hat{N}(\omega) \wedge N(\omega)} - \frac{2 \log N(\omega)}{\omega^2} - \sigma^2$$

$$\leq \frac{2}{\omega^2} \frac{\frac{L}{\sqrt{n}}}{\left(N(\omega) - \frac{L}{\sqrt{n}} \right)_+} - \frac{2 \log N(\omega)}{\omega^2} - \sigma^2 \tag{34}$$

where we have used the inequality $|\log(x) - \log(y)| \leq \frac{|x-y|}{x \wedge y}$ for $x, y > 0$. On the other hand, consider that for every $\omega \in \mathbb{R}$, we have on the event $\mathcal{E}$

$$-\frac{2 \log \hat{N}(\omega)}{\omega^2} \geq -\frac{2 \log \left(N(\omega) + \frac{L}{\sqrt{n}} \right)}{\omega^2}$$

$$= -\frac{2 \log \left(e^{-\frac{\sigma^2 \omega^2}{2}} \left| \frac{1}{n} \sum_{j=1}^n e^{i\omega(\theta+\gamma_j)} \right| + \frac{L}{\sqrt{n}} \right)}{\omega^2}$$

$$\geq -\frac{2 \log \left(e^{-\frac{\sigma^2 \omega^2}{2}} + \frac{L}{\sqrt{n}} \right)}{\omega^2}$$

$$= -\frac{2}{\omega^2} \left(\log \left(e^{-\frac{\sigma^2 \omega^2}{2}} + \frac{L}{\sqrt{n}} \right) - \log \left(e^{-\frac{\sigma^2 \omega^2}{2}} \right) \right) + \sigma^2.$$

Rearranging and using the inequality $|\log(x) - \log(y)| \leq \frac{|x-y|}{x \wedge y}$ for $x, y > 0$, we have

$$-\frac{2 \log \hat{N}(\omega)}{\omega^2} - \sigma^2 \geq -\frac{2}{\omega^2} \frac{\frac{L}{\sqrt{n}}}{e^{-\frac{\sigma^2 \omega^2}{2}}}.$$

Taking infimum over $\omega \in [a, b]$ yields

$$\hat{\sigma}^2 - \sigma^2 \geq - \sup_{\omega \in [a,b]} \frac{2}{\omega^2} \cdot \frac{L e^{\frac{\sigma^2 \omega^2}{2}}}{\sqrt{n}}. \tag{35}$$

To summarize, combining our upper and lower bounds yields

$$|\hat{\sigma}^2 - \sigma^2| \leq \left(\sup_{\omega \in [a,b]} \frac{2}{\omega^2} \cdot \frac{L e^{\frac{\sigma^2 \omega^2}{2}}}{\sqrt{n}} \right) + \inf_{\omega \in [a,b]} \left(\frac{2}{\omega^2} \frac{\frac{L}{\sqrt{n}}}{\left(N(\omega) - \frac{L}{\sqrt{n}} \right)_+} - \frac{2 \log N(\omega)}{\omega^2} - \sigma^2 \right)$$

$$= \left(\sup_{\omega \in [a,b]} \frac{2}{\omega^2} \cdot \frac{L e^{\frac{\sigma^2 \omega^2}{2}}}{\sqrt{n}} \right) + \inf_{\omega \in [a,b]} \left(\frac{2}{\omega^2} \frac{\frac{L}{\sqrt{n}}}{\left(N(\omega) - \frac{L}{\sqrt{n}} \right)_+} - \frac{2}{\omega^2} \log \left| \frac{1}{n} \sum_{j=1}^n e^{i\omega(\theta+\gamma_j)} \right| \right). \tag{36}$$

Define the event $\mathcal{E}' = \left\{ \frac{1}{L'} \leq \frac{\tilde{\sigma}}{\sigma} \leq L' \right\}$ where L' depends only on δ and, according to Proposition 3.1, ensures $\inf_{\substack{\theta \in \mathbb{R}, \\ ||\gamma||_0 \leq k, \\ \sigma > 0}} P_{\theta,\gamma,\sigma}(\mathcal{E}') \geq 1 - \frac{\delta}{2}$. Let us examine the second term in (36). Let $\mathcal{O}'$ be the set

given by $\mathcal{O}$ along with some arbitrary additional indices from $\mathcal{I}$ to ensure $|\mathcal{O}'| = k$. Then

$$\left| \frac{1}{n} \sum_{j=1}^{n} e^{i\omega(\theta + \gamma_j)} \right| = \left| 1 - \frac{k}{n} \left(1 - \frac{1}{k} \sum_{j \in \mathcal{O}'} e^{i\omega\gamma_j} \right) \right| \geq 1 - \frac{k}{n} \left| 1 - \frac{1}{k} \sum_{j \in \mathcal{O}'} e^{i\omega\gamma_j} \right|. \tag{37}$$

Consider that

$$\left| 1 - \frac{1}{k} \sum_{j \in \mathcal{O}'} e^{i\omega\gamma_j} \right|^2 = 1 + \left| \frac{1}{k} \sum_{j \in \mathcal{O}'} e^{i\omega\gamma_j} \right|^2 - \frac{2}{k} \sum_{j \in \mathcal{O}'} \cos(\omega\gamma_j) \leq 2 \left(1 - \frac{1}{k} \sum_{j \in \mathcal{O}'} \cos(\omega\gamma_j) \right).$$

Proposition 3.3 implies there exists $\omega^* \in [a,b]$ such that $\frac{1}{k} \sum_{j \in \mathcal{O}'} \cos(\omega^* \gamma_j) \geq -\frac{1}{5}$, and so (37) can be bounded as $\left| \frac{1}{n} \sum_{j=1}^{n} e^{i\omega^*(\theta + \gamma_j)} \right| \geq \frac{n - k\sqrt{\frac{12}{5}}}{n}$. Note this directly implies $N(\omega^*) \geq e^{-\frac{\sigma^2 (\omega^*)^2}{2}} \left(\frac{n - k\sqrt{\frac{12}{5}}}{n} \right)$. Therefore, the bound (36) reduces to

$$|\hat{\sigma}^2 - \sigma^2| \leq \left(\sup_{\omega \in [a,b]} \frac{2}{\omega^2} \cdot \frac{L e^{\frac{\sigma^2 \omega^2}{2}}}{\sqrt{n}} \right) + \frac{2}{(\omega^*)^2} \left(\frac{\frac{L}{\sqrt{n}}}{\left(e^{-\frac{\sigma^2 (\omega^*)^2}{2}} \left(1 - \sqrt{\frac{3}{5}} \right) - \frac{L}{\sqrt{n}} \right)_+} + \log \left(\frac{1}{1 - \frac{k}{n}\sqrt{\frac{12}{5}}} \right) \right)$$

$$\leq \frac{2}{a^2} \cdot \frac{L e^{\frac{\sigma^2 b^2}{2}}}{\sqrt{n}} + \frac{2}{a^2} \left(\frac{\frac{L}{\sqrt{n}}}{\left(e^{-\frac{\sigma^2 b^2}{2}} \left(1 - \sqrt{\frac{3}{5}} \right) - \frac{L}{\sqrt{n}} \right)_+} + \frac{4k}{n} \right) \tag{38}$$

where we have used $\frac{k}{n} < \frac{1}{2}$ to obtain $\frac{n - k\sqrt{\frac{12}{5}}}{n} \geq 1 - \sqrt{\frac{3}{5}}$. We have also used the inequality $\log\left(1/\left(1 - x\sqrt{\frac{12}{5}} \right) \right) \leq 4x$ for $0 \leq x \leq \frac{1}{2}$. Consider that $a = c\tilde{\sigma}^{-1} \sqrt{1 \vee \log\left(\frac{ek^2}{n} \right)}$. On the event $\mathcal{E}'$ we have $a \geq \frac{c}{L'} \sigma^{-1} \sqrt{1 \vee \log\left(\frac{ek^2}{n} \right)}$ and $b \leq 100 L' c \sigma^{-1} \sqrt{1 \vee \log\left(\frac{ek^2}{n} \right)}$. Choosing c appropriately small depending only on L, L' and noting n is sufficiently large depending only on δ, it thus follows from (38) that on the event $\mathcal{E} \cap \mathcal{E}'$

$$|\hat{\sigma}^2 - \sigma^2| \leq \frac{C' \sigma^2}{1 \vee \log\left(\frac{ek^2}{n} \right)} \left(\frac{1 \vee \frac{k}{\sqrt{n}}}{\sqrt{n}} + \frac{4k}{n} \right) \leq \frac{C' \sigma^2 k}{n \log\left(1 + \frac{k}{\sqrt{n}} \right)}$$

for some constant $C' > 0$ depending only on δ (whose value may change from instance to instance). Since union bound asserts $\mathcal{E} \cap \mathcal{E}'$ has $P_{\theta,\gamma,\sigma}$-probability at least $1 - \delta$ uniformly over the parameter space, the proof is complete. $\qquad\square$

References

[1] Antoniadis, A. (2007). Wavelet methods in statistics: some recent developments and their applications. *Stat. Surv.* 1:16–55.

[2] Azriel, D. and Schwartzman, A. (2015). The empirical distribution of a large number of correlated normal variables. *J. Amer. Statist. Assoc.* 110(511):1217–1228.

[3] Bahmani, S. (2021). Nearly optimal robust mean estimation via empirical characteristic function. *Bernoulli* 27(3):2139–2158.

[4] Belomestny, D. and Reiß, M. (2006). Spectral calibration of exponential Lévy models. *Finance Stoch.* 10(4):449–474.

[5] Belomestny, D. and Reiß, M. (2015). Estimation and Calibration of Lévy Models via Fourier Methods. In D. Belomestny, F. Comte, V. Genon-Catalot, H. Masuda, and M. Reiß, editors, *Lévy Matters IV: Estimation for Discretely Observed Lévy Processes*, Lecture Notes in Mathematics, pp. 1–76. Springer International Publishing, Cham.

[6] Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration inequalities.* Oxford University Press, Oxford.

[7] Cai, T. T. and Jin, J. (2010). Optimal rates of convergence for estimating the null density and proportion of nonnull effects in large-scale multiple testing. *Ann. Statist.* 38(1):100–145.

[8] Cai, T. T. and Sun, W. (2009). Simultaneous testing of grouped hypotheses: finding needles in multiple haystacks. *J. Amer. Statist. Assoc.* 104(488):1467–1481.

[9] Carpentier, A., Delattre, S., Roquain, E., and Verzelen, N. (2021). Estimating minimum effect with outlier selection. *Ann. Statist.* 49(1):272–294.

[10] Carpentier, A. and Verzelen, N. (2019). Adaptive estimation of the sparsity in the Gaussian vector model. *Ann. Statist.* 47(1):93–126.

[11] Carroll, R. J. and Hall, P. (1988). Optimal rates of convergence for deconvolving a density. *J. Amer. Statist. Assoc.* 83(404):1184–1186.

[12] Chen, M., Gao, C., and Ren, Z. (2016). A general decision theory for Huber's ϵ-contamination model. *Electron. J. Stat.* 10(2):3752–3774.

[13] Chen, M., Gao, C., and Ren, Z. (2018). Robust covariance and scatter matrix estimation under Huber's contamination model. *Ann. Statist.* 46(5):1932–1960.

[14] Chernoff, H. (1964). Estimation of the mode. *Ann. Inst. Statist. Math.* 16:31–41.

[15] Collier, O., Comminges, L., and Tsybakov, A. B. (2017). Minimax estimation of linear and quadratic functionals on sparsity classes. *Ann. Statist.* 45(3):923–958.

[16] Collier, O. and Dalalyan, A. S. (2019). Multidimensional linear functional estimation in sparse Gaussian models and robust estimation of the mean. *Electron. J. Stat.* 13(2):2830–2864.

[17] Comminges, L., Collier, O., Ndaoud, M., and Tsybakov, A. B. (2021). Adaptive robust estimation in sparse vector model. *Ann. Statist.* 49(3):1347–1377.

[18] Dalalyan, A. and Thompson, P. (2019). Outlier-robust estimation of a sparse linear model using ℓ_1-penalized Huber's M-estimator. In *Advances in Neural Information Processing Systems.*

[19] Devroye, L., Mehrabian, A., and Reddad, T. (2023). The total variation distance between high-dimensional Gaussians with the same mean. ArXiv:1810.08693 [math, stat].

[20] Diakonikolas, I., Kamath, G., Kane, D., Li, J., Moitra, A., and Stewart, A. (2019). Robust estimators in high-dimensions without the computational intractability. *SIAM J. Comput.* 48(2):742–864.

[21] Donoho, D. and Montanari, A. (2016). High dimensional robust M-estimation: asymptotic variance via approximate message passing. *Probab. Theory Related Fields* 166(3):935–969.

[22] Eddy, W. F. (1980). Optimum kernel estimators of the mode. *Ann. Statist.* 8(4):870–882.

[23] Efron, B. (2004). Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *J. Amer. Statist. Assoc.* 99(465):96–104.

[24] Efron, B. (2007). Correlation and large-scale simultaneous significance testing. *J. Amer. Statist. Assoc.* 102(477):93–103.

[25] Efron, B. (2008). Microarrays, empirical Bayes and the two-groups model. *Statist. Sci.* 23(1):1–22.

[26] Fan, J. (1991). On the optimal rates of convergence for nonparametric deconvolution problems. *Ann. Statist.* 19(3):1257–1272.

[27] Fan, J. and Han, X. (2017). Estimation of the false discovery proportion with unknown dependence. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* 79(4):1143–1164.

[28] Fan, J., Han, X., and Gu, W. (2012). Estimating false discovery proportion under arbitrary covariance dependence. *J. Amer. Statist. Assoc.* 107(499):1019–1035.

[29] Foygel, R. and Mackey, L. (2014). Corrupted sensing: novel guarantees for separating structured signals. *IEEE Trans. Inform. Theory* 60(2):1223–1247.

[30] Gannaz, I. (2007). Robust estimation and wavelet thresholding in partially linear models. *Stat. Comput.* 17(4):293–310.

[31] Heller, R. and Rosset, S. (2021). Optimal control of false discovery criteria in the two-group model. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* 83(1):133–155.

[32] Huber, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* 35:73–101.

[33] Indritz, J. (1961). An inequality for Hermite polynomials. *Proc. Amer. Math. Soc.* 12:981–983.

[34] Ingster, Y. I. and Suslina, I. A. (2003). *Nonparametric Goodness-of-Fit Testing Under Gaussian Models*, volume 169 of *Lecture Notes in Statistics*. Springer-Verlag, New York.

[35] Jin, J. (2008). Proportion of non-zero normal means: universal oracle equivalences and uniformly consistent estimators. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* 70(3):461–493.

[36] Jin, J. and Cai, T. T. (2007). Estimating the null and the proportional of nonnull effects in large-scale multiple comparisons. *J. Amer. Statist. Assoc.* 102(478):495–506.

[37] Kotekal, S. and Gao, C. (2023). Minimax rates for sparse signal detection under correlation. *Inf. Inference* 12(4):iaad044.

[38] Kotekal, S. and Gao, C. (2023). Sparsity meets correlation in Gaussian sequence model. ArXiv:2312.09356 [math, stat].

[39] Laurent, B. and Massart, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.* 28(5):1302–1338.

[40] Leek, J. T. and Storey, J. D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLOS Genetics* 3(9):e161.

[41] Leek, J. T. and Storey, J. D. (2008). A general framework for multiple testing dependence. *Proceedings of the National Academy of Sciences* 105(48):18718–18723.

[42] Lepski, O. V. (1990). A problem of adaptive estimation in Gaussian white noise. *Teor. Veroyatnost. i Primenen.* 35(3):459–470.

[43] Lepski, O. V. (1991). Asymptotically minimax adaptive estimation. I. Upper bounds. Optimally adaptive estimates. *Teor. Veroyatnost. i Primenen.* 36(4):645–659.

[44] Lepski, O. V. (1992). Asymptotically minimax adaptive estimation. II. Schemes without optimal adaptation. Adaptive estimates. *Teor. Veroyatnost. i Primenen.* 37(3):468–481.

[45] McCann, L. and Welsch, R. E. (2007). Robust variable selection using least angle regression and elemental set sampling. *Comput. Statist. Data Anal.* 52(1):249–257.

[46] Meister, A. (2009). *Deconvolution problems in nonparametric statistics*, volume 193 of *Lecture Notes in Statistics*. Springer-Verlag, Berlin.

[47] Nguyen, N. H. and Tran, T. D. (2013). Robust Lasso with missing and grossly corrupted observations. *IEEE Trans. Inform. Theory* 59(4):2036–2058.

[48] Nie, Y. and Wu, Y. (2023). Large-scale multiple testing: fundamental limits of false discovery rate control and compound oracle. ArXiv: 2302.06809 [math, stat].

[49] Owen, A. B. (2005). Variance of the number of false discoveries. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* 67(3):411–426.

[50] Parzen, E. (1962). On estimation of a probability density function and mode. *Ann. Math. Statist.* 33:1065–1076.

[51] Qiu, X., Brooks, A. I., Klebanov, L., and Yakovlev, N. (2005). The effects of normalization on the correlation structure of microarray data. *BMC Bioinformatics* 6(1):1–11.

[52] Qiu, X., Klebanov, L., and Yakovlev, A. (2005). Correlation between gene expression levels and limitations of the empirical bayes methodology for finding differentially expressed genes. *Stat Appl Genet Mol Biol* 4(1).

[53] Sardy, S., Tseng, P., and Bruce, A. (2001). Robust wavelet denoising. *IEEE Transactions on Signal Processing* 49(6):1146–1152. Conference Name: IEEE Transactions on Signal Processing.

[54] Schwartzman, A. (2010). Comment: "Correlated z-values and the accuracy of large-scale statistical estimates". *J. Amer. Statist. Assoc.* 105(491):1059–1063.

[55] She, Y. and Owen, A. B. (2011). Outlier detection using nonconvex penalized regression. *J. Amer. Statist. Assoc.* 106(494):626–639.

[56] Stephens, M. (2017). False discovery rates: a new deal. *Biostatistics* 18(2):275–294.

[57] Sun, L. and Stephens, M. (2018). Solving the empirical Bayes normal means problem with correlated noise. ArXiv: 1812.07488 [stat].

[58] Sun, W. and Cai, T. T. (2007). Oracle and adaptive compound decision rules for false discovery rate control. *J. Amer. Statist. Assoc.* 102(479):901–912.

[59] Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, New York.

[60] Vershynin, R. (2018). *High-dimensional Probability*, volume 47 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.

[61] Verzelen, N. (2012). Minimax risks for sparse regressions: ultra-high dimensional phenomenons. *Electron. J. Stat.* 6:38–90.

A   Fourier-based estimator: unknown variance

Proof of Theorem 3.4. Fix $\delta \in (0,1)$. It follows from an argument using the bounded differences inequality (e.g. Theorem D.13) that for a sufficiently large $L > 0$ depending only on δ, the event

$$\mathcal{E} := \left\{ \sup_{\omega \in \mathbb{R}} \left| \frac{1}{n} \sum_{j=1}^n e^{i\omega(X_j - \theta)} - E_{\theta,\gamma,\sigma}\big(e^{i\omega(X_j - \theta)}\big) \right| \leq \frac{L}{\sqrt{n}} \right\}$$

has $P_{\theta,\gamma,\sigma}$-probability of at least $1 - \frac{\delta}{2}$ uniformly over the parameter space. Let us also define the event

$$\mathcal{E}_{\mathrm{var}} := \left\{ \sigma^2 \in [\sigma_-^2, \sigma_+^2] \right\}.$$

For an appropriate R, Theorem 3.2 asserts that $\mathcal{E}_{\mathrm{var}}$ has $P_{\theta,\gamma,\sigma}$-probability of at least $1 - \frac{\delta}{2}$ uniformly over the parameter space. For ease of notation, let us define the function

$$F(\mu, v) := \sup_{|\omega| \leq \tau} \inf_{\substack{\zeta \in \mathbb{C}, \\ |\zeta| \leq 1}} \left| \frac{1}{n} \sum_{j=1}^n e^{i\omega(X_j - \mu) + \frac{v^2 \omega^2}{2}} - \frac{n-k}{n} - \frac{k}{n} \zeta \right|.$$

By definition of $\hat{\theta}$ and $\hat{v}$, we have on the event $\mathcal{E} \cap \mathcal{E}_{\mathrm{var}}$

$$F(\hat{\theta}, \hat{v}) \leq F(\theta, \sigma^2) \leq \frac{L e^{\frac{\sigma^2 \tau^2}{2}}}{\sqrt{n}} + \sup_{|\omega| \leq \tau} \inf_{\substack{\zeta \in \mathbb{C}, \\ |\zeta| \leq 1}} \left| \frac{1}{n} \sum_{j=1}^n e^{i\omega\gamma_j} - \frac{n-k}{n} - \frac{k}{n} \zeta \right| = \frac{L e^{\frac{\sigma^2 \tau^2}{2}}}{\sqrt{n}}. \tag{39}$$

On the other hand, consider that on the event $\mathcal{E} \cap \mathcal{E}_{\text{var}}$

$$
\begin{aligned}
F(\hat{\theta}, \hat{v}) &= \sup_{|\omega| \leq \tau} \inf_{\substack{\zeta \in \mathbb{C}, \\ |\zeta| \leq 1}} \left| \frac{1}{n} \sum_{j=1}^{n} e^{i\omega(X_j - \hat{\theta}) + \frac{\hat{v}\omega^2}{2}} - \frac{n-k}{n} - \frac{k}{n}\zeta \right| \\
&\geq -\frac{L e^{\frac{\hat{v}\tau^2}{2}}}{\sqrt{n}} + \sup_{|\omega| \leq \tau} \inf_{\substack{\zeta \in \mathbb{C}, \\ |\zeta| \leq 1}} \left| \frac{1}{n} \sum_{j=1}^{n} e^{i\omega(\theta + \gamma_j - \hat{\theta}) + \frac{(\hat{v} - \sigma^2)\omega^2}{2}} - \frac{n-k}{n} - \frac{k}{n}\zeta \right| \\
&= -\frac{L e^{\frac{\hat{v}\tau^2}{2}}}{\sqrt{n}} + \sup_{|\omega| \leq \tau} \inf_{\substack{\zeta \in \mathbb{C}, \\ |\zeta| \leq 1}} \left| \frac{n-k}{n}\left(e^{i\omega(\theta - \hat{\theta}) + \frac{(\hat{v} - \sigma^2)\omega^2}{2}} - 1\right) + \frac{k}{n}\left(\frac{1}{k}\sum_{j \in \mathcal{O}'} e^{i\omega(\gamma_j - \hat{\theta}) + \frac{(\hat{v} - \sigma^2)\omega^2}{2}} - \zeta\right) \right| \\
&\geq -\frac{L e^{\frac{\hat{v}\tau^2}{2}}}{\sqrt{n}} + \sup_{|\omega| \leq \tau} \inf_{\substack{z \in \mathbb{C}, \\ |z| \leq \alpha}} \left| \frac{n-k}{n}\left(e^{i\omega(\theta - \hat{\theta}) + \frac{(\hat{v} - \sigma^2)\omega^2}{2}} - 1\right) + \frac{k}{n}z \right| \qquad (40)
\end{aligned}
$$

where $\mathcal{O}'$ is the set obtained by taking $\mathcal{O}$ and adding arbitrary indices from $\mathcal{I}$ if needed to ensure $|\mathcal{O}'| = k$. Also, here we denote $\alpha = 1 + e^{\frac{(\hat{v} - \sigma^2)\omega^2}{2}}$. Similar to the proof of Theorem 2.2, we break up the analysis into two cases to separately obtain the parametric part and the nonparametric part of our target bound. Let $C_1 > 0$ be a sufficiently large constant depending only on δ.

Case 1: Suppose $k \leq C_1\sqrt{n}$. From (39) and (40) we have

$$
\sup_{|\omega| \leq \tau} \left| e^{i\omega(\theta - \hat{\theta}) + \frac{(\hat{v} - \sigma^2)\omega^2}{2}} - 1 \right| \leq \frac{n}{n-k}\left(\frac{k}{n}\left(1 + e^{\frac{|\hat{v} - \sigma^2|\tau^2}{2}}\right) + \frac{L\left(e^{\frac{\sigma^2\tau^2}{2}} + e^{\frac{\hat{v}^2\tau^2}{2}}\right)}{\sqrt{n}} \right).
$$

Since we are working on the event $\mathcal{E} \cap \mathcal{E}_{\text{var}}$ where we have $\hat{v}, \sigma^2 \in [\sigma_-^2, \sigma_+^2]$, by taking c sufficiently small depending only on δ we have

$$
|\hat{v} - \sigma^2|\tau^2 \leq \frac{2Rk}{n \log\left(1 + \frac{k}{\sqrt{n}}\right)} \cdot \left(1 \vee c^2 \log\left(1 + \frac{k^2(n-2k)^2}{n^3}\right)\right) \leq \frac{2Rk}{n \log\left(1 + \frac{k}{C_1\sqrt{n}}\right)} \leq \frac{4RC_1}{\sqrt{n}}
$$

where we have used that one can take $C_1 \geq 1$ and we have used the inequality $\frac{x}{2} \leq \log(1 + x)$ for $0 \leq x \leq 1$. Since n is sufficiently large to ensure $\frac{4RC_1}{\sqrt{n}}$, we have $1 + e^{\frac{|\hat{v} - \sigma^2|\tau^2}{2}} \leq 2 + \frac{2eRC_1}{\sqrt{n}}$ from the inequality $e^x \leq 1 + ex$ for $0 \leq x \leq 1$. Further, since n is sufficiently large and c sufficiently small we have

$$
\sigma^2\tau^2 = \frac{\sigma^2}{\hat{\sigma}^2}\left(1 \vee c^2 \log\left(1 + \frac{k^2(n-2k)^2}{n^3}\right)\right) \leq \frac{1}{1 - \frac{Rk}{n \log\left(1 + \frac{k}{\sqrt{n}}\right)}}\left(1 \vee c^2 \log\left(1 + \frac{k^2(n-2k)^2}{n^3}\right)\right) \leq 2,
$$

$$
\hat{v}^2\tau^2 \leq \left(1 + \frac{Rk}{n \log\left(1 + \frac{k}{\sqrt{n}}\right)}\right)\left(1 \vee c^2 \log\left(1 + \frac{k^2(n-2k)^2}{n^3}\right)\right) \leq 2.
$$

Putting together these bounds, we have

$$
\sup_{|\omega| \leq \tau} \left| e^{i\omega(\theta - \hat{\theta}) + \frac{(\hat{v} - \sigma^2)\omega^2}{2}} - 1 \right| \leq \frac{n}{n-k}\left(\frac{k}{n}\left(2 + \frac{2eRC_1}{\sqrt{n}}\right) + \frac{2Le}{\sqrt{n}}\right) \leq \frac{\tilde{C}}{\sqrt{n}}
$$

for some positive constant $\tilde{C}$ depending only on δ. We claim $|\hat{\theta} - \theta| \leq \frac{\tilde{C}'\hat{\sigma}}{\sqrt{n}}$ where $\tilde{C}'$ is a sufficiently large constant depending only on δ. To prove the claim, suppose not, that is, suppose $|\hat{\theta} - \theta| > \frac{\tilde{C}'\hat{\sigma}}{\sqrt{n}}$. Consider the choice $\omega^* = \frac{\tilde{C}'}{\sqrt{n}(\theta - \hat{\theta})}$ satisfies $|\omega^*| \leq \hat{\sigma}^{-1} \leq \tau$. Using the inequality $|ye^{ix} - 1| \geq |\cos(x)e^{ix} - 1|$ for all $x, y \in \mathbb{R}$, we have

$$
\left| \cos\left(\frac{\tilde{C}'}{\sqrt{n}}\right) e^{\frac{i\tilde{C}'}{\sqrt{n}}} - 1 \right| \leq \left| e^{i\omega^*(\theta - \hat{\theta}) + \frac{(\hat{v} - \sigma^2)(\omega^*)^2}{2}} - 1 \right| \leq \sup_{|\omega| \leq \tau} \left| e^{i\omega(\theta - \hat{\theta}) + \frac{(\hat{v} - \sigma^2)\omega^2}{2}} - 1 \right| \leq \frac{\tilde{C}}{\sqrt{n}}.
$$

Clearly this gives a contradiction when $\tilde{C}'$ and n are sufficiently large depending only on δ. Hence, we have established the claim $|\hat{\theta} - \theta| \leq \frac{\tilde{C}'\hat{\sigma}}{\sqrt{n}}$. Since we are working on the event $\mathcal{E} \cap \mathcal{E}_{\mathrm{var}}$, it thus follows $|\hat{\theta} - \theta| \leq \tilde{C}'' \frac{\sigma}{\sqrt{n}}$ for some constant $\tilde{C}''$ depending only on δ. Thus we have the desired result.

Case 2: Suppose $k > C_1 \sqrt{n}$. Consider the second term in (40). Analogous to the proof of Theorem 2.2, the minimizer solving

$$\inf_{\substack{z \in \mathbb{C}, \\ |z| \leq \alpha}} \left| \frac{n-k}{n} \left(e^{i\omega(\theta - \hat{\theta}) + \frac{(\hat{v} - \sigma^2)\omega^2}{2}} - 1 \right) + \frac{k}{n} z \right|$$

is

$$z^* = \begin{cases} -\frac{n-k}{k} \left(e^{i\omega(\theta - \hat{\theta}) + \frac{(\hat{v} - \sigma^2)\omega^2}{2}} - 1 \right) & \text{if } \left| e^{i\omega(\theta - \hat{\theta}) + \frac{(\hat{v} - \sigma^2)\omega^2}{2}} - 1 \right| \leq \frac{\alpha k}{n-k}, \\ -\frac{\alpha}{\left| e^{i\omega(\theta - \hat{\theta}) + \frac{(\hat{v} - \sigma^2)\omega^2}{2}} - 1 \right|} \left(e^{i\omega(\theta - \hat{\theta}) + \frac{(\hat{v} - \sigma^2)\omega^2}{2}} - 1 \right) & \text{otherwise.} \end{cases}$$

Therefore, arguing analogously as in the proof of Theorem 2.2, we have

$$\sup_{|\omega| \leq \tau} \inf_{\substack{z \in \mathbb{C}, \\ |z| \leq \alpha}} \left| \frac{n-k}{n} \left(e^{i\omega(\theta - \hat{\theta}) + \frac{(\hat{v} - \sigma^2)\omega^2}{2}} - 1 \right) + \frac{k}{n} z \right|$$

$$= \sup_{|\omega| \leq \tau} \left| \frac{n-k}{n} \left(e^{i\omega(\theta - \hat{\theta}) + \frac{(\hat{v} - \sigma^2)\omega^2}{2}} - 1 \right) + \frac{k}{n} z^* \right|$$

$$= \sup_{\substack{|\omega| \leq \tau, \\ \left| e^{i\omega(\theta - \hat{\theta}) + \frac{(\hat{v} - \sigma^2)\omega^2}{2}} - 1 \right| > \frac{\alpha k}{n-k}}} \left| \frac{n-k}{n} \left(e^{i\omega(\theta - \hat{\theta}) + \frac{(\hat{v} - \sigma^2)\omega^2}{2}} - 1 \right) + \frac{k}{n} z^* \right|$$

$$= \sup_{\substack{|\omega| \leq \tau, \\ \left| e^{i\omega(\theta - \hat{\theta}) + \frac{(\hat{v} - \sigma^2)\omega^2}{2}} - 1 \right| > \frac{\alpha k}{n-k}}} \left| \frac{n-k}{n} - \frac{\alpha k}{n \left| e^{i\omega(\theta - \hat{\theta}) + \frac{(\hat{v} - \sigma^2)\omega^2}{2}} - 1 \right|} \right| \left| e^{i\omega(\theta - \hat{\theta}) + \frac{(\hat{v} - \sigma^2)\omega^2}{2}} - 1 \right|.$$

To summarize, from (39) and (40), we have

$$\sup_{\substack{|\omega| \leq \tau, \\ \left| e^{i\omega(\theta - \hat{\theta}) + \frac{(\hat{v} - \sigma^2)\omega^2}{2}} - 1 \right| > \frac{\alpha k}{n-k}}} \left| \frac{n-k}{n} - \frac{\alpha k}{n \left| e^{i\omega(\theta - \hat{\theta}) + \frac{(\hat{v} - \sigma^2)\omega^2}{2}} - 1 \right|} \right| \left| e^{i\omega(\theta - \hat{\theta}) + \frac{(\hat{v} - \sigma^2)\omega^2}{2}} - 1 \right| \leq \frac{L \left(e^{\frac{\sigma^2 \tau^2}{2}} + e^{\frac{\hat{v}^2 \tau^2}{2}} \right)}{\sqrt{n}}.$$

(41)

The argument continues similarly as in the proof of Theorem 2.2. We claim (41) implies $|\hat{\theta} - \theta| \leq \frac{16k}{n} \tau^{-1}$. For sake of contradiction, suppose $|\hat{\theta} - \theta| > \frac{16k}{n} \tau^{-1}$. Consider the choice $\omega^* := \frac{\pi \wedge (16k/n)}{\theta - \hat{\theta}}$. Note $|\omega^*| = \frac{\pi \wedge (16k/n)}{|\hat{\theta} - \theta|} \leq \tau$. Let α^* denote the value of α corresponding to ω^*. If $\pi \leq 16k/n$, then $\omega^* = \frac{\pi}{\theta - \hat{\theta}}$ and so

$$\left| e^{i\omega^*(\theta - \hat{\theta}) + \frac{(\hat{v} - \sigma^2)(\omega^*)^2}{2}} - 1 \right| = \alpha^* > \frac{\alpha^* k}{n-k}.$$

Likewise, if $\pi > 16k/n$, then

$$\left| e^{i\omega^*(\theta - \hat{\theta}) + \frac{(\hat{v} - \sigma^2)(\omega^*)^2}{2}} - 1 \right| = \left| e^{i(16k/n) + \frac{(\hat{v} - \sigma^2)(\omega^*)^2}{2}} - 1 \right| > \frac{5k}{n}.$$

Here, we have used the inequality $|ye^{ix} - 1| > \frac{5}{16} x$ which holds at least on the region $\{(x, y) \in \mathbb{R}^2 : y > 0 \text{ and } 0 \leq x \leq \pi\}$. We have applied this inequality with the choice $x = \frac{16k}{n} < \pi$ and $y = \exp \left(\frac{(\hat{v} - \sigma^2)(\omega^*)^2}{2} \right) > 0$. Further consider that since $(\omega^*)^2 \leq \tau^2$ and since $\hat{v}, \sigma^2 \in [\sigma_-^2, \sigma_+^2]$ as we

are working on $\mathcal{E} \cap \mathcal{E}_{\mathrm{var}}$, we have

$$
\begin{aligned}
\alpha^* &= 1 + e^{\frac{(\hat{v} - \sigma^2)(\omega^*)^2}{2}} \\
&\leq 1 + e^{\frac{|\hat{v} - \sigma^2| \tau^2}{2}} \\
&\leq 1 + \exp\left(\frac{R}{2} \frac{k}{n \log\left(1 + \frac{k}{\sqrt{n}}\right)} \left(1 \vee c^2 \log\left(1 + \frac{k^2(n-2k)^2}{n^3}\right)\right) \right) \\
&\leq 1 + \exp\left(\frac{c^2 R k}{n} \right)
\end{aligned}
$$

where we have used $C_1 < k < \frac{n}{2} - C\sqrt{n}$ for C_1 and C sufficiently large. Consider that for c such that $c^2 R \leq 1$, we have $e^{c^2 R k/n} \leq 1 + \frac{ec^2 Rk}{n}$ by the inequality $e^x \leq 1 + ex$ for $x \in [0, 1]$. Therefore, taking c sufficiently small to ensure $ec^2 R \leq 1$ gives us $\alpha^* \leq 2 + \frac{k}{n} \leq \frac{5}{2}$. Hence,

$$
\left| e^{i\omega^*(\theta - \hat{\theta}) + \frac{(\hat{v} - \sigma^2)(\omega^*)^2}{2}} - 1 \right| > \frac{5k}{n} \geq \frac{\alpha^* k}{n - k}.
$$

In summary, ω^* lives in the domain of optimization in (41). Therefore, from (39)

$$
\begin{aligned}
\frac{L\left(e^{\frac{\sigma^2 \tau^2}{2}} + e^{\frac{\hat{v}^2 \tau^2}{2}}\right)}{\sqrt{n}} &\geq
\begin{cases}
\left(\frac{n-2k}{n}\right) \alpha^* & \text{if } \pi \leq 16k/n, \\
\left(\frac{n-k}{n} - \frac{\alpha^*}{5}\right)\left(\frac{5k}{n}\right) & \text{if } \pi > 16k/n
\end{cases} \\
&\geq
\begin{cases}
\frac{4 - \pi}{8}\left(\frac{n-2k}{n}\right) & \text{if } \pi \leq 16k/n, \\
\frac{4 - \pi}{8}\left(\frac{5k}{n}\right) & \text{if } \pi > 16k/n
\end{cases} \\
&\geq
\begin{cases}
\frac{4 - \pi}{8}\left(\frac{n-2k}{n}\right)\left(\frac{k}{n}\right) & \text{if } \pi \leq 16k/n, \\
\frac{4 - \pi}{8}\left(\frac{k}{n}\right)\left(\frac{n-2k}{n}\right) & \text{if } \pi > 16k/n
\end{cases} \\
&= \frac{4 - \pi}{8} \cdot \frac{(n - 2k)k}{n^2}.
\end{aligned}
$$

Note we have used $\alpha^* \geq 1$. Rearranging, we have $1 \leq \frac{16L}{4 - \pi} \cdot \frac{n^{3/2}}{(n-2k)k}\left(e^{\frac{\sigma^2 \tau^2}{2}} + e^{\frac{\hat{v}^2 \tau^2}{2}}\right)$. Since $\tau^2 = \hat{\sigma}^{-2}\left(1 \vee c^2 \log\left(1 + \frac{k^2(n-2k)^2}{n^3}\right)\right)$, since n is larger than a sufficiently large constant depending only on δ, and since $C_1 < k < \frac{n}{2} - C\sqrt{n}$ for C_1 and C sufficiently large, we have

$$
\begin{aligned}
\sigma^2 \tau^2 &= \frac{\sigma^2}{\hat{\sigma}^2}\left(1 \vee c^2 \log\left(1 + \frac{k^2(n-2k)^2}{n^3}\right)\right) \\
&\leq \frac{1}{1 - \frac{Rk}{n \log\left(1 + \frac{k}{\sqrt{n}}\right)}}\left(1 \vee c^2 \log\left(1 + \frac{k^2(n-2k)^2}{n^3}\right)\right) \\
&\leq 2c^2 \log\left(1 + \frac{k^2(n-2k)^2}{n^3}\right),
\end{aligned}
$$

and

$$
\hat{v}^2 \tau^2 \leq \left(1 + \frac{Rk}{n \log\left(1 + \frac{k}{\sqrt{n}}\right)}\right)\left(1 \vee c^2 \log\left(1 + \frac{k^2(n-2k)^2}{n^3}\right)\right) \leq 2c^2 \log\left(1 + \frac{k^2(n-2k)^2}{n^3}\right).
$$

Taking c sufficiently small and C sufficiently large yields

$$
1 \leq \frac{32L}{4 - \pi} \cdot \frac{n^{3/2}}{(n-2k)k} e^{2c^2 \log\left(1 + \frac{k^2(n-2k)^2}{n^3}\right)} < 1
$$

which is a contradiction. Hence, on the event $\mathcal{E} \cap \mathcal{E}_{\mathrm{var}}$ we have $|\hat{\theta} - \theta| \leq \frac{32k}{n}\tau^{-1}$, which clearly yields our desired result. $\square$

B   Lower bounds

In this section, we prove Theorems 4.1 and 5.2. Recall that the Fourier transform of an integrable function f is $\hat{f}(t) = \int e^{-itx} f(x)\, dx$ and the Fourier transform of a finite Borel measure π is $\hat{\pi}(t) = \int e^{-itx} \pi(dx)$. Recall also that $P_{\theta, \gamma}$ denotes the joint distribution of the data $\{X_i\}_{i=1}^n$ generated from the model (1) with parameters θ and γ (recall from Section 2 that suppression of σ in the notation denotes $\sigma = 1$).

B.1   Proof of Theorem 4.1

Proof of Theorem 4.1. For ease of notation, set

$$\psi = \frac{k}{n} \log^{-1/2} \left(1 + \frac{k^2 (n-2k)^2}{n^3} \right).$$

If n is bounded above by a universal constant, it is trivial to show the lower bound $\psi \asymp 1$. Suppose n is larger than a sufficiently big universal constant. If $k \lesssim \sqrt{n}$, then $\psi \asymp \frac{k}{n} \log^{-1/2} \left(1 + \frac{k^2}{n} \right) \asymp \frac{k}{n} \cdot \frac{\sqrt{n}}{k} = n^{-1/2}$ where we have used $\frac{x}{2} \leq \log(1+x) \leq x$ for $x \in [0,1]$. Thus ψ is the parametric rate and a very standard two point argument will establish the lower bound. Hence, we can limit our attention to the case $k \geq C' \sqrt{n}$ for a sufficiently large universal $C' > 0$ and n larger than a sufficiently big universal constant.

We now construct the priors for (θ, γ) to establish the lower bound. Let f_0 and f_1 denote the probability densities given by Proposition B.2 with the choice $\varepsilon = \frac{k - 10\sqrt{n}}{n}$ and $c = 1$. Let μ, τ, λ, g_0, and g_1 be the associated quantities from Proposition B.2. Note that

$$\mu = \frac{c_0 \lambda}{\tau} = \left(\frac{c_0}{1 + 2\varepsilon} \right) \frac{\varepsilon}{1 \vee B_c \sqrt{\log(en\varepsilon^2 (1 - 2\varepsilon)^2)}} \asymp \psi$$

where c_0 is a small universal constant and B_c is a large universal constant. Note to conclude $\mu \asymp \psi$ we have used that $k \leq \frac{n}{2} - \sqrt{n}$ implies $1 - 2\varepsilon = \frac{n - 2k + 20\sqrt{n}}{n} \asymp \frac{n-2k}{n}$. Define the priors π_0 and π_1 as follows. Let $\delta_1, ..., \delta_n \overset{iid}{\sim} \text{Bernoulli}(\varepsilon)$. A draw $(\theta, \gamma) \sim \pi_0$ is obtained by setting $\theta = 0$ and setting $\gamma = (\delta_1 v_1, ..., \delta_n v_n)$ where $v_1, ..., v_n \overset{iid}{\sim} g_0 * \delta_\mu$. Likewise, a draw $(\theta, \gamma) \sim \pi_1$ is obtained by setting $\theta = 2\mu$ and setting $\gamma = (\delta_1 v_1, ..., \delta_n v_n)$ where $(v_1, ..., v_n) \overset{iid}{\sim} g_1 * \delta_{-\mu}$.

Note that π_0 and π_1 are not supported on the parameter space. Namely, it is not the case that $||\gamma||_0 \leq k$ with probability one under either prior since the δ_i are independent. We are only guaranteed $E\left(||\gamma||_0\right) \leq k$. Consequently, define the truncated versions $\bar{\pi}_0$ and $\bar{\pi}_1$ where for any event A,

$$\bar{\pi}_i(A) = \frac{\pi_i \left(A \cap \{ ||\gamma||_0 \leq k \} \right)}{\pi_i \left(\{ ||\gamma||_0 \leq k \} \right)}$$

for $i = 0, 1$. Denote the mixture distributions $P_i = \int P_{\theta, \gamma} \pi_i(d\theta, d\gamma)$ and $\bar{P}_i = \int P_{\theta, \gamma} \bar{\pi}_i(d\theta, d\gamma)$ for $i = 0, 1$. Note that P_0 and P_1 admits densities $f_0^{\otimes n}$ and $f_1^{\otimes n}$ where f_0 and f_1 are given in Proposition B.2. With these definitions in hand, consider

$$\inf_{\hat{\theta}} \sup_{\substack{\theta \in \mathbb{R}, \\ ||\gamma||_0 \leq k}} P_{\theta, \gamma} \left\{ |\hat{\theta} - \theta| \geq \mu \right\}$$

$$\geq \inf_{\hat{\theta}} \max \left\{ \bar{P}_0 \left\{ |\hat{\theta} - \theta| \geq \mu \right\}, \bar{P}_1 \left\{ |\hat{\theta} - \theta| \geq \mu \right\} \right\}$$

$$= \inf_{\hat{\theta}} \max \left\{ \bar{P}_0 \left\{ |\hat{\theta}| \geq \mu \right\}, \bar{P}_1 \left\{ |\hat{\theta} - 2\mu| \geq \mu \right\} \right\}$$

$$\geq \inf_{\hat{\theta}} \max \left\{ \bar{P}_0 \left\{ |\hat{\theta}| > \mu \right\}, \bar{P}_1 \left\{ |\hat{\theta}| \leq \mu \right\} \right\}$$

$$\geq \inf_{A} \max \left\{ \bar{P}_0(A), \bar{P}_1(A^c) \right\}$$

$$\geq \frac{1}{2} \left(1 - d_{\text{TV}}(\bar{P}_1, \bar{P}_0) \right)$$

where the infimum runs over all events A. We have also used $\left\{|\hat{\theta} - 2\mu| \geq \mu\right\} \supset \left\{|\hat{\theta}| \leq \mu\right\}$, which follows from triangle inequality. Note triangle inequality further gives $d_{\mathrm{TV}}(\bar{P}_1, \bar{P}_0) \leq d_{\mathrm{TV}}(P_1, P_0) + d_{\mathrm{TV}}(P_0, \bar{P}_0) + d_{\mathrm{TV}}(\bar{P}_1, P_1)$. The data processing inequality gives $d_{\mathrm{TV}}(\bar{P}_i, P_i) \leq d_{\mathrm{TV}}(\bar{\pi}_i, \pi_i)$ for $i = 0, 1$. Consider that for any event B, we have $\pi_i(B) - \bar{\pi}_i(B) \leq \pi_i(B) - \pi_i(B \cap \{||\gamma||_0 \leq k\}) \leq \pi_i(\{||\gamma||_0 > k\})$. Likewise, $\pi_i(B) - \bar{\pi}_i(B) = \pi_i(B \cap \{||\gamma||_0 \leq k\}) + \pi_i(B \cap \{||\gamma||_0 > k\}) - \frac{\pi_i(B \cap \{||\gamma||_0 \leq k\})}{\pi_i(\{||\gamma||_0 \leq k\})} \geq \pi_i(B \cap \{||\gamma||_0 \leq k\}) \left(1 - \frac{1}{\pi_i(\{||\gamma||_0 \leq k\})}\right) = -\pi_i(\{||\gamma||_0 > k\})$. Therefore, we have $|\pi_i(B) - \bar{\pi}_i(B)| \leq \pi_i(\{||\gamma||_0 > k\})$. Taking supremum over all events B yields

$$d_{\mathrm{TV}}(\pi_i, \bar{\pi}_i) = \pi_i(\{||\gamma||_0 > k\}) \leq P\left\{\sum_{i=1}^{n} \delta_i > k\right\} = P\left\{\mathrm{Binomial}\,(n, \varepsilon) > k\right\} \leq \frac{k - 10\sqrt{n}}{100n} \leq \frac{1}{100}$$

where we have used Chebyshev's inequality. To summarize, these calculations imply $d_{\mathrm{TV}}(\bar{P}_1, \bar{P}_0) \leq d_{\mathrm{TV}}(P_0, P_1) + \frac{1}{50}$ and so it remains to bound $d_{\mathrm{TV}}(P_0, P_1)$. Note that if $X \sim P_i$ for $i = 0, 1$, then $\{X_j\}_{j=1}^n$ are mutually independent. Therefore, by Proposition B.2 we have

$$d_{\mathrm{TV}}(P_1, P_0) \leq \frac{1}{2}\sqrt{\chi^2(P_1 \| P_0)} = \frac{1}{2}\sqrt{(1 + \chi^2(f_1 \| f_0))^n - 1} \leq \frac{1}{2}\sqrt{e - 1}.$$

Therefore, we have $d_{\mathrm{TV}}(\bar{P}_1, \bar{P}_0) \leq \frac{1}{2}\sqrt{e - 1} + \frac{1}{50}$. Thus,

$$\inf_{\hat{\theta}} \sup_{\substack{\theta \in \mathbb{R}, \\ ||\gamma||_0 \leq k}} P_{\theta, \gamma}\left\{|\hat{\theta} - \theta| \geq \mu\right\} \geq \frac{1}{2}\left(1 - \frac{1}{2}\sqrt{e - 1} + \frac{1}{50}\right) \geq \frac{1}{10}.$$

Since $\mu \asymp \psi$, we have the desired result. The proof is complete. $\qquad\square$

Proposition B.1. *Suppose $0 < \lambda \leq \frac{1}{2}$ and $\tau > 0$. There exists a universal positive constant $c_0 < 1$ such that the following holds. Let $\mu = \frac{c_0\lambda}{\tau}$ and define the probability density function $p_0 : \mathbb{R} \to [0, \infty)$ with*

$$p_0(x) = \begin{cases} \frac{\tau}{4} & \text{if } |x| \leq \frac{1}{\tau}, \\ \frac{1}{4\tau x^2} & \text{if } |x| > \frac{1}{\tau}. \end{cases}$$

Further define the real-valued function $\Delta = -\frac{\hat{h}}{2\pi}$ where $h : \mathbb{R} \to \mathbb{C}$ is the purely imaginary-valued function

$$h(t) = \begin{cases} 0 & \text{if } t < -2\tau, \\ -j(-t) & \text{if } -2\tau \leq t \leq -\tau, \\ k(t) & \text{if } -\tau < t < \tau, \\ j(t) & \text{if } \tau \leq t \leq 2\tau, \\ 0 & \text{if } t > 2\tau, \end{cases}$$

with $k(t) = 2i\frac{1-\lambda}{\lambda}\sin(t\mu)$ and $j(t) = 2i\frac{1-\lambda}{\lambda}\sin(\tau\mu)\frac{2\tau - t}{\tau}$. Then the function $p_1 = p_0 + \Delta$ is a probability density function.

Proof. Since $p_1 = p_0 + \Delta$, it suffices to show p_1 is nonnegative and g integrates to zero in order to show p_1 is a probability density function. We first show p_1 is nonnegative provided c_0 is sufficiently small. Since h is a purely-imaginary odd function, it follows that

$$\Delta(x) = -\frac{\hat{h}(x)}{2\pi} = -\frac{1}{2\pi}\left(-2i\int_0^{\tau} k(t)\sin(tx)\,dt - 2i\int_{\tau}^{2\tau} j(t)\sin(tx)\,dt\right).$$

By the power series representation $k(t) = 2i\frac{1-\lambda}{\lambda}\sin(t\mu) = 2i\frac{1-\lambda}{\lambda}\sum_{n=0}^{\infty}\frac{(-1)^n(t\mu)^{2n+1}}{(2n+1)!}$ and Lemma D.7,

$$-2i\int_0^\tau k(t)\sin(tx)\,dt$$
$$= 4\frac{1-\lambda}{\lambda}\sum_{n=0}^{\infty}\frac{(-1)^n(c_0\lambda)^{2n+1}}{(2n+1)!}\int_0^\tau\frac{t^{2n+1}}{\tau^{2n+1}}\sin(tx)\,dt$$
$$= 4\frac{1-\lambda}{\lambda}\sum_{n=0}^{\infty}\frac{(-1)^n(c_0\lambda)^{2n+1}}{(2n+1)!}\left(-\frac{\cos(\tau x)}{x}+R_{k,n}(x)\right).$$

Likewise, since $\tau\mu = c_0\lambda$ we have $j(t) = 2i\frac{1-\lambda}{\lambda}\sum_{n=0}^{\infty}\frac{(-1)^n(c_0\lambda)^{2n+1}}{(2n+1)!}\frac{2\tau-t}{\tau}$. By Lemma D.8

$$-2i\int_\tau^{2\tau} j(t)\sin(tx)\,dt = 4\frac{1-\lambda}{\lambda}\sum_{n=0}^{\infty}\frac{(-1)^n(c_0\lambda)^{2n+1}}{(2n+1)!}\left(\frac{\cos(\tau x)}{x}+R_j(x)\right).$$

Note that $|R_{k,n}(x)|\vee|R_j(x)| \le \frac{C(1\vee n)}{\tau x^2}\vee\frac{2}{\tau x^2}$ for a universal positive constant C by Lemmas D.7 and D.8. Therefore, it follows that

$$\Delta(x) = -\frac{4}{2\pi}\frac{1-\lambda}{\lambda}\sum_{n=0}^{\infty}\frac{(-1)^n(c_0\lambda)^{2n+1}}{(2n+1)!}(R_{k,n}(x)+R_j(x))$$

and so $|\Delta(x)| \le \frac{Cc_0}{\tau x^2}$ for a large universal constant $C > 0$. Therefore, selecting c_0 sufficiently small depending only on C ensures $|\Delta(x)| \le p_0(x)$ for $|x| > \frac{1}{\tau}$ and so p_1 is nonnegative here. Now let us show p_1 is nonnegative on $|x| \le \frac{1}{\tau}$. First, consider $I := \int_0^\tau\frac{t}{\tau}\sin(tx)\,dt$ satisfies $|I| \le \frac{1}{\tau}\int_0^\tau t\,dt = \frac{\tau}{2}$. Hence, $|-\frac{4}{2\pi}(1-\lambda)c_0 I| \le 2c_0\tau$. Further, consider that

$$\left|\Delta(x)-\left(-\frac{4}{2\pi}(1-\lambda)c_0 I\right)\right|$$
$$\le \frac{1}{2\pi}\left|-2i\int_0^\tau k(t)\sin(tx)\,dt - 4(1-\lambda)c_0\int_0^\tau\frac{t}{\tau}\sin(tx)\,dt\right| + \frac{1}{2\pi}\left|-2i\int_\tau^{2\tau} j(t)\sin(tx)\,dt\right|$$
$$\le \left|4\frac{1-\lambda}{\lambda}\int_0^\tau(\sin(t\mu)-t\mu)\sin(tx)\,dt\right| + \left|4\frac{1-\lambda}{\lambda}\sum_{n=0}^{\infty}\frac{(-1)^n(c_0\lambda)^{2n+1}}{(2n+1)!}\int_\tau^{2\tau}\frac{2\tau-t}{\tau}\sin(tx)\,dt\right|$$

where we have used $\tau = \frac{c_0\lambda}{\mu}$ to obtain the first term. Examining the first term, consider that $|t\mu| \le c_0\lambda$ for $|t| \le \tau$ since $\mu = \frac{c_0\lambda}{\tau}$. Therefore, $|\sin(t\mu)-t\mu| \le c_0^3\lambda^3/6$. Thus,

$$\left|4\frac{1-\lambda}{\lambda}\int_0^\tau(\sin(t\mu)-t\mu)\sin(tx)\,dt\right| \le 4\frac{1-\lambda}{\lambda}\int_0^\tau|\sin(t\mu)-t\mu|\,dt \le \frac{4}{6}c_0\tau.$$

Here, we have used $c_0^3 \le c_0$ (as we can take $c_0 < 1$) and $\lambda^2(1-\lambda) \le 1$ just to simplify expressions. Examining the second term, provided $c_0 < 1$ we have

$$4\frac{1-\lambda}{\lambda}\left|\sum_{n=0}^{\infty}\frac{(-1)^n(c_0\lambda)^{2n+1}}{(2n+1)!}\int_\tau^{2\tau}\frac{2\tau-t}{\tau}\sin(tx)\,dt\right| \le 4\frac{1-\lambda}{\lambda}\sum_{n=0}^{\infty}\frac{(c_0\lambda)^{2n+1}}{(2n+1)!}\cdot\frac{1}{\tau}\int_\tau^{2\tau}|2\tau-t|\,dt$$
$$\le 4\frac{1-\lambda}{\lambda}\sum_{n=0}^{\infty}\frac{(c_0\lambda)^{2n+1}}{(2n+1)!}\cdot\frac{\tau}{2}$$
$$\le 2c_0\tau\sum_{n=0}^{\infty}\frac{1}{(2n+1)!}$$
$$= 2c_0\sinh(1)\tau.$$

Therefore, $\left|\Delta(x)-\left(-\frac{4}{2\pi}(1-\lambda)c_0 I\right)\right| \le \frac{4}{6}c_0\tau + 2c_0\sinh(1)\tau \le 4c_0\tau$. Therefore, from the bound $\left|-\frac{4}{2\pi}(1-\lambda)c_0 I\right| \le 2c_0\tau$, it follows that $\Delta \ge -6c_0\tau \ge -\frac{\tau}{4}$ provided we have taken $c_0 \le \frac{1}{24}$. Thus, it immediately follows that p_1 is nonnegative for $|x| \le \frac{1}{\tau}$, which concludes the proof of the claim that p_1 is a nonnegative function.

It remains to show p_1 integrates to one, and to do so it suffices to show $\int\Delta = 0$. Note Δ is

uniformly bounded and $|\Delta(x)| \leq \frac{Cc_0}{\tau x^2}$ as established earlier. Hence, Δ, and also $\hat{h}$, is integrable. Therefore, $\hat{\Delta}(t) = -\frac{1}{2\pi} \int \hat{h}(x)e^{-itx}\, dx = -h(-t) = h(t)$ since h is an odd function. Therefore, $\int \Delta(x)\, dx = \hat{\Delta}(0) = h(0) = 0$. Thus p_1 is a probability density function as claimed. $\qquad\blacksquare$

Proposition B.2. *Suppose $0 < \varepsilon \leq \frac{1}{2}$ and $n \in \mathbb{N}$. For any $c > 0$, there exists a large $B_c > 0$ depending only on c such that the following holds. Let $\mu, p_0,$ and p_1 be given as in the statement of Proposition B.1 with the choice $\lambda = \frac{\varepsilon}{1+2\varepsilon}$ and $\tau = 1 \vee B_c\sqrt{\log\left(en\varepsilon^2(1-2\varepsilon)^2\right)}$. Define the probability measures $g_1 = 2\varepsilon\delta_{-\mu} + (1-2\varepsilon)p_1$ and $g_0 = 2\varepsilon\delta_\mu + (1-2\varepsilon)p_0$ and the probability densities*

$$f_0 = ((1-\varepsilon)\delta_0 + \varepsilon(g_0 * \delta_\mu)) * \varphi,$$
$$f_1 = ((1-\varepsilon)\delta_{2\mu} + \varepsilon(g_1 * \delta_\mu)) * \varphi$$

where φ denotes the probability density function of the standard Gaussian distribution and $*$ denotes convolution. Then $\chi^2(f_1 \,\|\, f_0) \leq \frac{c}{n}$.

Proof. Note that $f_0 \geq (1-\varepsilon)\varphi$. Since $\varphi^{-1}(x) = \sqrt{2\pi}\sum_{k=0}^{\infty} \frac{x^{2k}}{2^k k!}$, we have

$$\chi^2(f_1 \,\|\, f_0) = \int \frac{(f_1-f_0)^2}{f_0}$$
$$\leq \frac{\sqrt{2\pi}}{1-\varepsilon} \sum_{k=0}^{\infty} \frac{1}{2^k k!} \int x^{2k}(f_1(x) - f_0(x))^2\, dx$$
$$= \frac{\sqrt{2\pi}}{1-\varepsilon} \sum_{k=0}^{\infty} \frac{1}{2^k k!} \int \left| \hat{f}_1^{(k)}(t) - \hat{f}_0^{(k)}(t) \right|^2 dt.$$

Since the Fourier transform of a convolution of densities is the product of the Fourier transforms of the densities, we have

$$\frac{\hat{f}_1(t) - \hat{f}_0(t)}{\hat{\varphi}(t)} = (1-\varepsilon)\left(e^{-2it\mu} - 1\right) + \varepsilon e^{-it\mu}\left(\hat{g}_1(t) - \hat{g}_0(t)\right)$$
$$= (1-\varepsilon)\left(e^{-2it\mu} - 1\right) + \varepsilon e^{-it\mu}\left(2\varepsilon(e^{it\mu} - e^{-it\mu}) + (1-2\varepsilon)\hat{\Delta}(t)\right)$$
$$= (1-\varepsilon - 2\varepsilon^2)(e^{-2it\mu} - 1) + \varepsilon(1-2\varepsilon)e^{-it\mu}\hat{\Delta}(t)$$
$$= (1-2\varepsilon)(1+\varepsilon)(e^{-2it\mu} - 1) + \varepsilon(1-2\varepsilon)e^{-it\mu}\hat{\Delta}(t)$$
$$= (1-2\varepsilon)(1+2\varepsilon)e^{-it\mu}\left((1-\lambda)(e^{-it\mu} - e^{it\mu}) + \lambda\hat{\Delta}(t)\right)$$

where $\Delta = p_1 - p_0$ is the function from Proposition B.1. Recall, as argued in Proposition B.1, we have $\hat{\Delta}(t) = h(t)$. Consequently, we have

$$\hat{f}_1(t) - \hat{f}_0(t) = \begin{cases} 0 & \text{if } |t| < \tau, \\ (1-2\varepsilon)(1+2\varepsilon)e^{-it\mu}\left((1-\lambda)(e^{-it\mu} - e^{it\mu}) + \lambda\hat{\Delta}(t)\right)\hat{\varphi}(t) & \text{if } \tau \leq |t| \leq 2\tau, \\ (1-2\varepsilon)(1+\varepsilon)(e^{-2it\mu} - 1)\hat{\varphi}(t) & \text{if } |t| > 2\tau. \end{cases}$$

Since $\hat{f}_1(t) - \hat{f}_0(t)$ vanishes for $|t| < \tau$, our bound for the χ^2 divergence simplifies

$$\chi^2(f_1 \,\|\, f_0) \leq \frac{\sqrt{2\pi}}{1-\varepsilon} \sum_{k=0}^{\infty} \frac{1}{2^k k!} \int_{|t| \geq \tau} \left| \hat{f}_1^{(k)}(t) - \hat{f}_0^{(k)}(t) \right|^2 dt. \tag{42}$$

The remainder of the proof is concerned with bounding this integral by splitting it across two different

regions. Let us first consider the region $|t| > 2\tau$. For such t, observe that

$$\left| \hat{f}_1^{(k)}(t) - \hat{f}_0^{(k)}(t) \right| = (1 - 2\varepsilon)(1 + \varepsilon) \left| \sum_{\ell=0}^{k} \binom{k}{\ell} \hat{\varphi}^{(\ell)}(t) \left(e^{-2it\mu} - 1 \right)^{(k-\ell)} \right|$$

$$= (1 - 2\varepsilon)(1 + \varepsilon) \left| (e^{-2it\mu} - 1)\hat{\varphi}^{(k)}(t) + \sum_{\ell=0}^{k-1} \binom{k}{\ell} \hat{\varphi}^{(\ell)}(t)(-2i\mu)^{k-\ell} e^{-2it\mu} \right|$$

$$\leq 2(1 - 2\varepsilon)(1 + \varepsilon)|\mu t| \left| \hat{\varphi}^{(k)}(t) \right| + 2(1 - 2\varepsilon)(1 + \varepsilon)|\mu| \sum_{\ell=0}^{k-1} \binom{k}{\ell} |\hat{\varphi}^{(\ell)}(t)|$$

where we have used $2|\mu| \leq 1$ (as c_0 in the definition of μ can be taken sufficiently small and we have $\tau \geq 1$) implies $|2\mu|^{k-\ell} \leq 2|\mu|$ for $k - \ell \geq 1$. Note also we have used $|e^{-2it\mu} - 1| \leq |2t\mu|$ to obtain the first term. Consider $\hat{\varphi}(t) = e^{-\frac{t^2}{2}}$ and that Corollary D.10 gives $|\hat{\varphi}^{(\ell)}(t)| \leq e^{-\frac{t^2}{4}}\sqrt{\ell!}$. Therefore,

$$\frac{\sqrt{2\pi}}{1 - \varepsilon} \sum_{k=0}^{\infty} \frac{1}{2^k k!} \int_{|t| \geq 2\tau} \left| \hat{f}_1^{(k)}(t) - \hat{f}_0^{(k)}(t) \right|^2 dt$$

$$\lesssim (1 - 2\varepsilon)^2 \mu^2 \frac{(1 + \varepsilon)^2}{1 - \varepsilon} \sum_{k=0}^{\infty} \frac{1}{2^k k!} \left(\int_{|t| \geq 2\tau} t^2 \left| \hat{\varphi}^{(k)}(t) \right|^2 dt + \int_{|t| \geq 2\tau} \left(\sum_{\ell=0}^{k-1} \binom{k}{\ell} |\hat{\varphi}^{(\ell)}(t)| \right)^2 dt \right)$$

$$\lesssim (1 - 2\varepsilon)^2 \mu^2 e^{-L\tau^2} + (1 - 2\varepsilon)^2 \mu^2 \sum_{k=0}^{\infty} \frac{1}{2^k k!} \left(\sum_{\ell=0}^{k-1} \binom{k}{\ell} \sqrt{\ell!} \right)^2 \int_{|t| \geq 2\tau} e^{-\frac{t^2}{2}} dt$$

$$\lesssim (1 - 2\varepsilon)^2 \mu^2 e^{-L\tau^2}$$

$$\lesssim (1 - 2\varepsilon)^2 \varepsilon^2 e^{-L\tau^2} \tag{43}$$

where $L > 0$ is a universal constant whose value may change from line to line. We have applied Lemma D.11 to obtain the penultimate line and $\mu^2 = \frac{c_0^2 \lambda^2}{\tau^2} \leq c_0^2 \varepsilon^2$ to obtain the final line. Let us now consider the region $\tau \leq |t| \leq 2\tau$. Recalling $\hat{\Delta}(t) = h(t)$ as argued at the end of Proposition B.1, on this interval we have

$$\left| \hat{f}_1^{(k)}(t) - \hat{f}_0^{(k)}(t) \right|^2 \leq 2(1 + 2\varepsilon)^2 (1 - 2\varepsilon)^2 (1 - \lambda)^2 \left| \frac{d^k}{dt^k} (e^{-2it\mu} - 1)\hat{\varphi}(t) \right|^2$$

$$+ 2(1 + 2\varepsilon)^2 (1 - 2\varepsilon)^2 \lambda^2 \left| \frac{d^k}{dt^k} e^{-it\mu} j(t) \hat{\varphi}(t) \right|^2 \tag{44}$$

where $j(t) = 2i\frac{1-\lambda}{\lambda}\sin(\tau\mu)\frac{2\tau - t}{\tau}$ is the function from Proposition B.1. By repeating an argument similar to that yielding (43), the first term in (44) yields the bound

$$\frac{\sqrt{2\pi}}{1 - \varepsilon} \sum_{k=0}^{\infty} \frac{1}{2^k k!} \int_{\tau \leq |t| \leq 2\tau} 2(1 + 2\varepsilon)^2 (1 - 2\varepsilon)^2 (1 - \lambda)^2 \left| \frac{d^k}{dt^k} (e^{-2it\mu} - 1)\hat{\varphi}(t) \right|^2 dt \lesssim (1 - 2\varepsilon)^2 \varepsilon^2 e^{-L\tau^2}. \tag{45}$$

It remains to bound the second term in (44). Consider that $j^{(k)}(t) = 0$ for all $k \geq 2$. Further consider $|j^{(k)}(t)| \leq C\frac{1-\lambda}{\lambda}|\sin(\tau\mu)| \leq \frac{C}{\lambda} \cdot |\tau\mu| = Cc_0$ for all $\tau \leq t \leq 2\tau$ and $k = 0, 1$. Here, $C > 0$ is a universal constant whose value may change from instance to instance. Therefore, for any $k \in \mathbb{N} \cup \{0\}$ we have

$$\left| \frac{d^k}{dt^k} e^{-it\mu} j(t) \hat{\varphi}(t) \right| = \left| \sum_{\ell=0}^{k} \binom{k}{\ell} \left(e^{-it\mu} j(t) \right)^{(\ell)} \hat{\varphi}^{(k-\ell)}(t) \right| \leq \sum_{\ell=0}^{k} \binom{k}{\ell} \left| \left(e^{-it\mu} j(t) \right)^{(\ell)} \right| \left| \hat{\varphi}^{(k-\ell)}(t) \right|.$$

Consider that

$$\left| \left(e^{-it\mu} j(t) \right)^{(\ell)} \right| = \left| \sum_{r=0}^{\ell} \binom{\ell}{r} \left(e^{-it\mu} \right)^{\ell-r} j^{(r)}(t) \right| \leq \sum_{r=0}^{\ell} \binom{\ell}{r} |\mu|^{\ell-r} |j^{(r)}(t)| \leq Cc_0 \max_{0 \leq r \leq 1} \binom{\ell}{r} \leq Cc_0(\ell \vee 1).$$

We have used $|\mu| \leq 1$. Thus, we have

$$
\left| \frac{d^k}{dt^k} e^{-it\mu} j(t)\hat{\varphi}(t) \right| \leq Cc_0 \sum_{\ell=0}^{k} \binom{k}{\ell} \left| \hat{\varphi}^{(k-\ell)}(t) \right| (\ell \vee 1)
$$

$$
\leq Cc_0(k \vee 1) \sum_{\ell=0}^{k} \binom{k}{\ell} \left| \hat{\varphi}^{(k-\ell)}(t) \right|
$$

$$
\leq Cc_0(k \vee 1) \sum_{\ell=0}^{k} \binom{k}{\ell} \left| \hat{\varphi}^{(\ell)}(t) \right|.
$$

Arguing similarly to before, we have via Corollary D.10

$$
\frac{\sqrt{2\pi}}{1-\varepsilon} \sum_{k=0}^{\infty} \frac{1}{2^k k!} \int_{\tau \leq |t| \leq 2\tau} 2(1+2\varepsilon)^2(1-2\varepsilon)^2 \lambda^2 \left| \frac{d^k}{dt^k} e^{-it\mu} j(t)\hat{\varphi}(t) \right|^2
$$

$$
\leq Cc_0^2 (1-2\varepsilon)^2 \lambda^2 \sum_{k=0}^{\infty} \frac{k^2 \vee 1}{2^k k!} \int_{\tau \leq |t| \leq 2\tau} \left(\sum_{\ell=0}^{k} \binom{k}{\ell} \left| \hat{\varphi}^{(\ell)}(t) \right| \right)^2 dt
$$

$$
\leq Cc_0^2 (1-2\varepsilon)^2 \lambda^2 \sum_{k=0}^{\infty} \frac{k^2 \vee 1}{2^k k!} \left(\sum_{\ell=0}^{k} \binom{k}{\ell} \sqrt{\ell!} \right)^2 \int_{\tau \leq |t| \leq 2\tau} e^{-\frac{t^2}{2}} dt
$$

$$
\leq Cc_0^2 (1-2\varepsilon)^2 \varepsilon^2 e^{-L\tau^2}. \tag{46}
$$

Here, we have used $\lambda^2 \leq \varepsilon^2$ and Lemma D.11. Plugging (45) and (46) into the corresponding integral of (44) yields

$$
\frac{\sqrt{2\pi}}{1-\varepsilon} \sum_{k=0}^{\infty} \frac{1}{2^k k!} \int_{\tau \leq |t| \leq 2\tau} \left| \hat{f}_1^{(k)}(t) - \hat{f}_0^{(k)}(t) \right|^2 dt \lesssim (1-2\varepsilon)^2 \varepsilon^2 e^{-L\tau^2} \tag{47}
$$

where we have used $\mu = \frac{c_0 \lambda}{\tau} \lesssim \varepsilon$. Combining (47) and (43) and taking B_c sufficiently large depending only on c, we have

$$
\chi^2 \left(f_1 \,\|\, f_0 \right) \leq C(1-2\varepsilon)^2 \varepsilon^2 e^{-L\tau^2} \leq C(1-2\varepsilon)^2 \varepsilon^2 e^{-LB_c^2 \log\left(en\varepsilon^2(1-2\varepsilon)^2\right)} \leq \frac{c}{n}.
$$

Note the choice of B_c depends only on c since C and L are some universal constants. The proof is complete. $\qquad\square$

B.2 Proof of Theorem 5.2

In this section, we give a proof of Theorem 5.2. Unlike in the proof of Theorem 4.1, we do not consider the related mixture formulation. Passing to the mixture formulation is only useful when $k \leq \frac{n}{2} - C\sqrt{n}$ since the number of outliers behaves like $k \pm O\left(\sqrt{n}\right)$ in the mixture formulation when $k \asymp n$. Instead, we work directly with the model (1). The lower bound construction is quite similar to that found in [38]. However, since the context of [38] is quite different, we give a full proof for completeness and for the reader's clarity.

Proof of Theorem 5.2. Recall we take $\sigma^2 = 1$ without loss of generality. For ease of notation, set $\psi^2 = \log\left(1 + \frac{n}{(n-2k)^2}\right)$. For further ease, let us assume without loss of generality $n/2$ is an integer. Define the sets $E_0 := \left\{1, ..., \frac{n}{2}\right\}$ and $E_1 := \left\{\frac{n}{2}+1, ..., n\right\}$. Let $C > 0$ be a suitably small constant. Let π_0 denote the prior in which θ is equal to $C\psi$ and $\gamma = -2C\psi \mathbf{1}_{S_0}$ where $S_0 \subset E_0$ is a subset of size k drawn uniformly at random. Likewise, let π_1 denote the prior in which $\theta = -C\psi$ and $\gamma = 2C\psi \mathbf{1}_{S_1}$ where $S_1 \subset E_1$ is a subset of size k drawn uniformly at random. Observe that if $(\theta_0, \gamma_0) \sim \pi_0$ and $(\theta_1, \gamma_1) \sim \pi_1$, then $|\theta_0 - \theta_1| = 2C\psi$. With this separation in hand, an invocation of Proposition D.3 yields

$$
\inf_{\hat{\theta}} \sup_{\substack{\theta \in \mathbb{R}, \\ ||\gamma||_0 \leq k}} P_{\theta,\gamma} \left\{ |\theta - \hat{\theta}| > C\psi \right\} \geq \frac{1}{2}(1 - \mathrm{d}_{\mathrm{TV}}(P_{\pi_0}, P_{\pi_1}))
$$

where $P_{\pi_j} = \int P_{\theta,\gamma} \pi_j(d\theta, d\gamma)$ for $j = 0, 1$ denotes the corresponding mixture measures. It remains to show the total variation is small to obtain the desired result.

Consider the random vector $X \in \mathbb{R}^n$ drawn from the mixture induced from the prior $X \sim P_{\pi_j}$ for $j = 0, 1$. Let us write $X_{E_0}, X_{E_1} \in \mathbb{R}^{n/2}$ to denote the random vectors of dimension $n/2$ obtained by taking the coordinates in E_0 and E_1 respectively. Note X_{E_0} and X_{E_1} are independent due to the construction of the priors. Let us write $P_{\pi_j}^I$ and $P_{\pi_j}^{II}$ to denote the marginal distributions of X_{E_0} and X_{E_1} respectively. By the independence, we have

$$d_{\mathrm{TV}}(P_{\pi_0}, P_{\pi_1}) \leq d_{\mathrm{TV}}(P_{\pi_0}^I, P_{\pi_1}^I) + d_{\mathrm{TV}}(P_{\pi_0}^{II}, P_{\pi_1}^{II}) = 2\, d_{\mathrm{TV}}(P_{\pi_0}^I, P_{\pi_1}^I). \tag{48}$$

The equality follows from the fact the marginal distributions $P_{\pi_0}^I$ and $P_{\pi_1}^{II}$, likewise $P_{\pi_0}^{II}$ and $P_{\pi_1}^I$, are the same up to a deterministic sign flip; hence the two total variation terms are equal in the above display.

It remains to bound $d_{\mathrm{TV}}(P_{\pi_0}^I, P_{\pi_1}^I)$. For ease of notation, set $Y = X_{E_0}$, $Q_0 = P_{\pi_0}^I$, and $Q_1 = P_{\pi_1}^I$. Observe that

$$Q_0 = \int N(-C\psi \mathbf{1}_{S_0} + C\psi \mathbf{1}_{S_0^c}, I_{n/2}) \pi_0(dS_0)$$
$$Q_1 = N(-C\psi \mathbf{1}_{\frac{n}{2}}, I_{n/2}).$$

By Neyman-Pearson lemma, the quantity $1 - d_{\mathrm{TV}}(P_{\pi_0}^I, P_{\pi_1}^I)$ is the optimal Type I plus Type II error for the testing problem

$$H_0 : Y \sim Q_0,$$
$$H_1 : Y \sim Q_1.$$

Equivalently, one can shift the data and consider the problem

$$H_0 : \tilde{Y} \sim \tilde{Q}_0,$$
$$H_1 : \tilde{Y} \sim \tilde{Q}_1$$

where $\tilde{Y} = Y + C\psi \mathbf{1}_{\frac{n}{2}}$ and $\tilde{Q}_0, \tilde{Q}_1$ are the corresponding distributions of $\tilde{Y}$ for when $Y \sim Q_0$ or $Y \sim Q_1$ respectively. Consider that

$$\tilde{Q}_0 = \int N(2C\psi \mathbf{1}_{S_0^c}, I_{n/2}) \pi_0(dS_0)$$
$$\tilde{Q}_1 = N(0, I_{n/2}).$$

In summary, $d_{\mathrm{TV}}(P_{\pi_0}^I, P_{\pi_1}^I) = d_{\mathrm{TV}}(\tilde{Q}_0, \tilde{Q}_1) \leq \frac{1}{2}\sqrt{\chi^2(\tilde{Q}_0 \,\|\, \tilde{Q}_1)}$. By the Ingster-Suslina method, we have

$$\chi^2(\tilde{Q}_0 \,\|\, \tilde{Q}_1) = E\left(\exp\left(\langle \mu, \mu' \rangle\right)\right) - 1 \tag{49}$$

where $\mu = 2C\psi \mathbf{1}_{S_0^c}$ and $\mu' = 2C\psi \mathbf{1}_{(S_0')^c}$ for S_0, S_0' independent and uniformly drawn from the collection of size k subsets of E_0. For ease of notation, denote $T_0 = S_0^c$ and $T_0' = (S_0')^c$. Noting that

$|T_0 \cap T_0'|$ is a hypergeometric random variable, Lemma D.5 yields

$$
\begin{aligned}
E\left(\exp\left(\langle\mu,\mu'\rangle\right)\right) &= E\left(\exp\left(4C^2\psi^2|T_0\cap T_0'|\right)\right) \\
&\leq \left(1 - \frac{n/2-k}{n/2} + \frac{n/2-k}{n/2}e^{4C^2\psi^2}\right)^{n/2-k} \\
&= \left(1 - \frac{n/2-k}{n/2} + \frac{n/2-k}{n/2}\left(1 + \frac{n}{(n-2k)^2}\right)^{4C^2}\right)^{n/2-k} \\
&\leq \left(1 + \frac{2C^2}{n/2-k}\right)^{n/2-k} \\
&\leq e^{2C^2}
\end{aligned}
$$

where we have used $2C^2 \leq 1$ (since C is taken sufficiently small) along with the inequality $(1+x)^\delta \leq 1 + \delta x$ for $0 \leq \delta \leq 1$ and $x \geq 0$. Therefore, from (49) we have $\chi^2(\tilde{Q}_0\,||\,\tilde{Q}_1) \leq e^{2C^2} - 1$, and so we can conclude $\mathrm{d}_{\mathrm{TV}}(P_{\pi_0}^I, P_{\pi_1}^I) \leq \frac{1}{2}\sqrt{e^{2C^2} - 1}$. From (48) it thus follows

$$
\inf_{\hat{\theta}} \sup_{\substack{\theta\in\mathbb{R},\\ ||\gamma||_0\leq k}} P_{\theta,\gamma}\left\{|\theta - \hat{\theta}| > C\psi\right\} \geq \frac{1}{2}(1 - \sqrt{e^{2C^2}-1}).
$$

Taking C sufficiently small, the right hand side can be bounded below by some positive universal constant c. The proof is complete. $\qquad\square$

C   Adaptation to k

Recall the setup of Section 5.3. The notation of Section 5.3 is freely used in the following.

Proof of Theorem 5.5. Define the event

$$
\mathcal{G} := \left\{L_\eta^{-1} \leq \frac{\sigma^2}{\tilde{\sigma}^2} \leq L_\eta\right\} \cap \bigcap_{\substack{k\in\mathcal{K},\\ k\geq k^*}} \left\{\frac{|\hat{\theta}_k - \theta|}{\sigma} \leq C'_{\delta,\eta}\epsilon(k,n)\right\}
$$

Consider that

$$
\begin{aligned}
&\sup_{\substack{\theta\in\mathbb{R},\\ ||\gamma||_0\leq k^*,\\ \sigma>0}} P_{\theta,\gamma,\sigma}\left\{\frac{|\hat{\theta}-\theta|}{\sigma} > 2C'_{\delta,\eta}L_\eta\epsilon(k^*,n)\right\} \\
&\leq \sup_{\substack{\theta\in\mathbb{R},\\ ||\gamma||_0\leq k^*,\\ \sigma>0}} P_{\theta,\gamma,\sigma}\left(\left\{\frac{|\hat{\theta}-\theta|}{\sigma} > 2C'_{\delta,\eta}L_\eta\epsilon(k^*,n)\right\}\cap\mathcal{G}\right) + \sup_{\substack{\theta\in\mathbb{R},\\ ||\gamma||_0\leq k^*,\\ \sigma>0}} P_{\theta,\gamma,\sigma}(\mathcal{G}^c). \quad (50)
\end{aligned}
$$

Consider that by Corollary C.1 we have $P_{\theta,\gamma,\sigma}(\mathcal{G}^c) \leq \eta + \delta$. It remains to bound the first term of (50). Consider that

$$
\mathcal{G} \subset \left\{L_\eta^{-1} \leq \frac{\sigma^2}{\tilde{\sigma}^2} \leq L_\eta\right\} \cap \bigcap_{\substack{k\in\mathcal{K},\\ k\geq k^*}} \left\{|\hat{\theta}_k - \theta| \leq \tilde{\sigma}C'_{\delta,\eta}\sqrt{L_\eta}\epsilon(k,n)\right\}.
$$

Consequently, on $\mathcal{G}$ it follows from the definition of k' in (28) that $k^* \geq k'$. Importantly, we thus have $\hat{\theta} \in J_{k^*}$, and so it immediately follows by triangle inequality

$$
\mathcal{G} \subset \left\{L_\eta^{-1} \leq \frac{\sigma^2}{\tilde{\sigma}^2} \leq L_\eta\right\} \cap \left\{|\hat{\theta}-\theta| \leq 2\tilde{\sigma}C'_{\delta,\eta}\sqrt{L_\eta}\epsilon(k^*,n)\right\} \subset \left\{|\hat{\theta}-\theta| \leq 2\sigma C'_{\delta,\eta}L_\eta\epsilon(k^*,n)\right\}.
$$

Therefore, the first term of (50) is zero. We have the desired result by taking $C_{\delta,\eta}$ sufficiently large. $\qquad\square$

Corollary C.1. *Fix $\delta \in (0,1)$. There exist constants $C', \tilde{C}, c > 0$ depending only on δ such that the*

following holds. If n is sufficiently large depending only on δ and $1 \leq k^* < \frac{n}{2} - \tilde{C}\sqrt{n}$, then

$$\sup_{\substack{\theta \in \mathbb{R}, \\ ||\gamma||_0 \leq k^*, \\ \sigma > 0}} P_{\theta,\gamma,\sigma} \left(\bigcup_{k^* \leq k < \frac{n}{2} - \tilde{C}\sqrt{n}} \left\{ \frac{|\hat{\theta}_k - \theta|}{\sigma} > \frac{C'k}{n\sqrt{\log\left(1 + \frac{k^2(n-2k)^2}{n^3}\right)}} \right\} \right) \leq \delta$$

where $\hat{\theta}_k$ is given by (22).

Proof. The proof is largely the same as the proof of Theorem 3.4. Let $\mathcal{E}$ be the event from the proof of Theorem 3.4 with the choice L depending on δ such that for any $\theta \in \mathbb{R}$, $||\gamma||_0 \leq k^*$, and $\sigma > 0$, we have $P_{\theta,\gamma,\sigma}(\mathcal{E}^c) \leq \delta/2$. For each k, let $\mathcal{E}_{\text{var},k} = \left\{\sigma^2 \in [\sigma^2_{-,k}, \sigma^2_{+,k}]\right\}$ where $\sigma^2_{-,k}, \sigma^2_{+,k}$ are the quantities associated to $\hat{\theta}_k$ in (22). We claim

$$P_{\theta,\gamma,\sigma}\left(\bigcap_{k^* \leq k < \frac{n}{2}} \mathcal{E}_{\text{var},k} \right) \geq 1 - \frac{\delta}{2}$$

uniformly over $\theta \in \mathbb{R}$, $||\gamma||_0 \leq k^*$, and $\sigma > 0$. As argued in Theorem 3.2, we have $\mathcal{E}_{\text{var},k} \supset \tilde{\mathcal{E}}$ where $\tilde{\mathcal{E}}$ is the event $\mathcal{E} \cap \mathcal{E}'$ in the proof of Theorem 3.2 (except now at confidence level $\delta/2$ instead of δ). Importantly, $\tilde{\mathcal{E}}$ does not depend on k and we have $P_{\theta,\gamma,\sigma}(\tilde{\mathcal{E}}) \geq 1 - \delta/2$ uniformly over $\theta \in \mathbb{R}$, $||\gamma||_0 \leq k^*$, and $\sigma > 0$. Thus, the claim is proved. Then, as argued in Theorem 3.4, we have

$$\mathcal{E} \cap \bigcap_{k^* \leq k < \frac{n}{2}} \mathcal{E}_{\text{var},k} \subset \bigcap_{k^* \leq k < \frac{n}{2} - \tilde{C}\sqrt{n}} \left\{ \frac{|\hat{\theta}_k - \theta|}{\sigma} \leq \frac{C'k}{n\sqrt{\log\left(1 + \frac{k^2(n-2k)^2}{n^3}\right)}} \right\}$$

since $\tilde{C}$ is sufficiently large. Thus, we have the desired result since the probability of the event on the left hand side is at least $1 - \delta$ by union bound. $\square$

The following results pertain to the construction of the variance estimator for the purposes of adaptive null estimation in Section 5.2. Recall the setup of Section 5.2 as we will use that notation freely in what follows.

Proof of Theorem 5.6. The proof proceeds in the same manner as the proof of Theorem 5.5. As in that argument, define the event

$$\mathcal{G} := \left\{ L_\eta^{-1} \leq \frac{\sigma^2}{\hat{\sigma}^2} \leq L_\eta \right\} \cap \bigcap_{k^* \leq k < \frac{n}{2}} \left\{ \frac{|\hat{\sigma}_k^2 - \sigma^2|}{\sigma^2} \leq \frac{C'_{\delta,\eta}k}{n\log\left(1 + \frac{k}{\sqrt{n}}\right)} \right\}.$$

Analogous to (50), we have

$$\sup_{\substack{\theta \in \mathbb{R}, \\ ||\gamma||_0 \leq k^*, \\ \sigma > 0}} P_{\theta,\gamma,\sigma} \left\{ \frac{|\hat{\sigma}^2 - \sigma^2|}{\sigma^2} > \frac{2C'_{\delta,\eta}L_\eta^2 k^*}{n\log\left(1 + \frac{k^*}{\sqrt{n}}\right)} \right\}$$

$$\leq \sup_{\substack{\theta \in \mathbb{R}, \\ ||\gamma||_0 \leq k^*, \\ \sigma > 0}} P_{\theta,\gamma,\sigma} \left(\left\{ \frac{|\hat{\sigma}^2 - \sigma^2|}{\sigma^2} > \frac{2C'_{\delta,\eta}L_\eta^2 k^*}{n\log\left(1 + \frac{k^*}{\sqrt{n}}\right)} \right\} \cap \mathcal{G} \right) + \sup_{\substack{\theta \in \mathbb{R}, \\ ||\gamma||_0 \leq k^*, \\ \sigma > 0}} P_{\theta,\gamma,\sigma}\left(\mathcal{G}^c\right). \qquad (51)$$

Consider that by Corollary C.2 we have $P_{\theta,\gamma,\sigma}(\mathcal{G}^c) \leq \eta + \delta$. To bound the first term in (51), consider that

$$\mathcal{G} \subset \left\{ L_\eta^{-1} \leq \frac{\sigma^2}{\hat{\sigma}^2} \leq L_\eta \right\} \cap \bigcap_{k^* \leq k < \frac{n}{2}} \left\{ |\hat{\sigma}_k^2 - \sigma^2| \leq \tilde{\sigma}^2 \frac{C'_{\delta,\eta}L_\eta k}{n\log\left(1 + \frac{k}{\sqrt{n}}\right)} \right\}.$$

Consequently, on $\mathcal{G}$ it follows from the definition of k' in (30) that $k^* \geq k'$ and so $\hat{\sigma}^2 \in J_{k^*}$. It

immediately follows by triangle inequality that

$$\mathcal{G} \subset \left\{ \frac{|\hat{\sigma}^2 - \sigma^2|}{\sigma^2} \leq \frac{2C'_{\delta,\eta} L_\eta^2 k^*}{n \log\left(1 + \frac{k^*}{\sqrt{n}}\right)} \right\}$$

and so the first term of (51) is zero. Thus, we have proved the desired result. $\qquad\square$

Corollary C.2. *Fix $\delta \in (0, 1)$. There exists $C > 0$ depending only on δ such that the following holds. If n is sufficiently large depending only on δ and $1 \leq k^* < \frac{n}{2}$, then*

$$\sup_{\substack{\theta \in \mathbb{R}, \\ ||\gamma||_0 \leq k^*, \\ \sigma > 0}} P_{\theta, \gamma, \sigma} \left(\bigcup_{k^* \leq k < \frac{n}{2}} \left\{ \frac{|\hat{\sigma}_k^2 - \sigma^2|}{\sigma^2} > \frac{Ck}{n \log\left(1 + \frac{k}{\sqrt{n}}\right)} \right\} \right) \leq \delta$$

where $\hat{\sigma}_k^2$ is the variance estimator from Theorem 3.2.

Proof. The proof is the same as the proof of Theorem 3.2, and the argument's structure is the same as that of the proof of Corollary C.1. In particular, let $\mathcal{E} \cap \mathcal{E}'$ be the event from the proof of Theorem 3.2, having made choices L, L' depending only on δ such that for any $\theta \in \mathbb{R}$, $||\gamma||_0 \leq k^*$, and $\sigma > 0$, we have $P_{\theta, \gamma, \sigma}(\mathcal{E}^c \cup \mathcal{E}'^c) \leq \delta$. The arguments of the proof of Theorem 3.2 show

$$\mathcal{E} \cap \mathcal{E}' \subset \left\{ \frac{|\hat{\sigma}_k^2 - \sigma^2|}{\sigma^2} \leq \frac{Ck}{n \log\left(1 + \frac{k}{\sqrt{n}}\right)} \right\}$$

for $k \geq k^*$. Since $\mathcal{E} \cap \mathcal{E}'$ does not depend on k, it follows

$$\mathcal{E} \cap \mathcal{E}' \subset \bigcap_{k^* \leq k < \frac{n}{2}} \left\{ \frac{|\hat{\sigma}_k^2 - \sigma^2|}{\sigma^2} \leq \frac{Ck}{n \log\left(1 + \frac{k}{\sqrt{n}}\right)} \right\},$$

and so we have the desired result. $\qquad\square$

D   Auxiliary results

Proof of Proposition 3.3. Let $p(x) = \lambda e^{-\lambda x} \mathbb{1}_{\{x > 0\}}$ denote the Exponential(λ) distribution with $\lambda = \frac{1}{50\alpha}$. Note this distribution has mean $\frac{1}{\lambda}$ and variance $\frac{1}{\lambda^2}$. Let $W \sim p$ and consider from Lemma D.6

$$E(f(W)) = \frac{1}{k} \sum_{j=1}^k \int \cos(\omega \gamma_j) p(\omega) d\omega = \frac{1}{k} \sum_{j=1}^k \frac{\lambda^2}{\lambda^2 + \gamma_j^2} \geq 0.$$

For ease of notation, denote $\beta = 100\alpha$. Let $\bar{p}$ denote the conditional density of W conditional on $W \in [\alpha, \beta]$. Note $\bar{p}$ is supported on $[\alpha, \beta]$. Then

$$\begin{aligned} \max_{\omega \in [\alpha, \beta]} f(\omega) &\geq \int f(\omega) \bar{p}(\omega) d\omega \\ &= \frac{E(f(W) \mathbb{1}_{\{W \in [\alpha, \beta]\}})}{P\{W \in [\alpha, \beta]\}} \\ &= \frac{E(f(W)) - E(f(W) \mathbb{1}_{\{W \in [\alpha, \beta]^c\}})}{P\{W \in [\alpha, \beta]\}} \\ &\geq -\frac{P\{W \in [\alpha, \beta]^c\}}{P\{W \in [\alpha, \beta]\}}. \end{aligned}$$

Consider that

$$P\{W \in [\alpha, \beta]^c\} = 1 - e^{-\lambda\alpha} + e^{-\lambda\beta} = 1 - e^{-\frac{1}{50}} + e^{-2} \leq 0.16$$

Therefore,

$$\max_{\omega \in [\alpha, \beta]} f(\omega) \geq -\frac{P\{W \in [\alpha, \beta]^c\}}{P\{W \in [\alpha, \beta]\}} \geq -\frac{0.16}{1 - 0.16} \geq -\frac{1}{5}.$$

The proof is complete. $\qquad\square$

Lemma D.1 (Lemma 1 [39])**.** *For any positive integer d and $t > 0$, we have*

$$P\left\{\chi_d^2 \geq d + 2\sqrt{dt} + 2t\right\} \leq e^{-t}.$$

Lemma D.2 (Lemma 11.1 [61])**.** *For any positive integer d and $t > 0$, we have*

$$P\left\{\chi_d^2 \leq e^{-1}dt^{2/d}\right\} \leq t.$$

Proposition D.3 (Method of two fuzzy hypotheses [59])**.** *Suppose $\mathcal{P}$ is a collection of distributions on a sample space $\mathcal{X}$ indexed by Θ and (Υ, ρ) is a metric space. Let $\tau : \Theta \to \Upsilon$ be a function, $\phi : \mathbb{R}_+ \to \mathbb{R}_+$ be a non-decreasing function with $\phi(0) = 0$. If $\Theta_0, \Theta_1 \subset \Theta$ and π_0, π_1 are two priors supported on Θ_0, Θ_1 respectively, then*

$$\inf_{\hat{\tau}} \sup_{\theta \in \Theta} P_\theta \left\{ \phi\left(\rho(\hat{\tau}(X), \tau(\theta))\right) > \phi(\delta) \right\} \geq \frac{1}{2}\left(1 - \mathrm{d_{TV}}(P_{\pi_0}, P_{\pi_1})\right)$$

where $2\delta := \inf_{\substack{\theta_0 \in \Theta_0 \\ \theta_1 \in \Theta_1}} \rho(\tau(\theta_0), \tau(\theta_1))$ and $P_{\pi_j} = \int P_\theta\, \pi_j$ for $j = 0, 1$.

Proposition D.4 (Ingster-Suslina method [34])**.** *Suppose $\Sigma \in \mathbb{R}^{n \times n}$ is a positive definite matrix and $\Theta \subset \mathbb{R}^n$ is a parameter space. Let P_θ denote the distribution $N(\theta, \Sigma)$. If π is a probability distribution supported on Θ, then*

$$\chi^2(P_\pi || P_0) = E\left(\exp\left(\langle \theta, \Sigma^{-1}\tilde{\theta}\rangle\right)\right) - 1$$

where $\theta, \tilde{\theta} \overset{iid}{\sim} \pi$. Here, $P_\pi = \int_\theta P_\theta \pi(d\theta)$ and $\chi^2(\cdot||\cdot)$ denotes the χ^2-divergence.

Lemma D.5 ([15])**.** *Suppose $1 \leq k \leq n$. If Y is distributed according to the hypergeometric distribution with probability mass function $P\{Y = \ell\} = \frac{\binom{k}{\ell}\binom{n-k}{k-\ell}}{\binom{n}{k}}$ for $0 \leq \ell \leq k$, then*

$$E(e^{\lambda Y}) \leq \left(1 - \frac{k}{n} + \frac{k}{n}e^\lambda\right)^k$$

for $\lambda > 0$.

Lemma D.6. *For $\lambda > 0$, define the probability density function $f(x) = \lambda e^{-\lambda x}\mathbb{1}_{\{x \geq 0\}}$ for the Exponential(λ) distribution. If $X \sim$ Exponential(λ), then*

$$E(\cos(tX)) = \frac{\lambda^2}{\lambda^2 + t^2}.$$

Proof. Consider that $E(\cos(tX))$ is the real part of the characteristic function $E(e^{itX}) = \frac{\lambda}{\lambda - it}$. Calculating directly,

$$\frac{\lambda}{\lambda - it} = \frac{\lambda}{\lambda + \frac{t^2}{\lambda}} + \frac{it\lambda}{\lambda^2 + t^2}.$$

Therefore,

$$E(\cos(tX)) = \frac{\lambda}{\lambda + \frac{t^2}{\lambda}} = \frac{\lambda^2}{\lambda^2 + t^2}.$$

$\qquad\square$

Lemma D.7. *If $\tau > 0$ and n is a nonnegative integer, then*

$$\int_0^\tau \frac{t^{2n+1}}{\tau^{2n+1}} \sin(tx)\, dt = -\frac{\cos(\tau x)}{x} + R(x)$$

with $|R(x)| \leq \frac{C(1 \vee n)}{\tau x^2}$ where $C > 0$ is a universal constant.

Proof. The proof follows by applying integration by parts twice. Suppose $n \geq 1$. Consider

$$\int_0^\tau \frac{t^{2n+1}}{\tau^{2n+1}} \sin(tx)\, dt = -\frac{\cos(\tau x)}{x} + \int_0^\tau \frac{(2n+1)t^{2n}}{\tau^{2n+1}} \cdot \frac{\cos(tx)}{x}\, dt$$
$$= -\frac{\cos(\tau x)}{x} + \frac{(2n+1)\sin(\tau x)}{\tau x^2} - \int_0^\tau \frac{2n(2n+1)t^{2n-1}}{\tau^{2n+1}} \cdot \frac{\sin(tx)}{x^2}\, dt.$$

Clearly $\left| \frac{(2n+1)\sin(\tau x)}{\tau x^2} \right| \leq \frac{Cn}{\tau x^2}$. Likewise, consider

$$\left| \int_0^\tau \frac{2n(2n+1)t^{2n-1}}{\tau^{2n+1}} \cdot \frac{\sin(tx)}{x^2}\, dt \right| \leq \frac{2n(2n+1)}{x^2} \int_0^\tau \frac{t^{2n-1}}{\tau^{2n+1}}\, dt \leq \frac{Cn}{\tau x^2}$$

and so we have the desired result. If $n = 0$, the same calculation gives $\int_0^\tau \frac{t}{\tau} \sin(tx)\, dt = -\frac{\cos(\tau x)}{x} + \frac{\sin(\tau x)}{\tau x^2}$, which yields the desired result. □

Lemma D.8. *If $\tau > 0$, then*

$$\int_\tau^{2\tau} \frac{2\tau - t}{\tau} \sin(tx)\, dt = \frac{\cos(\tau x)}{x} + R(x)$$

where $|R(x)| \leq \frac{2}{\tau x^2}$.

Proof. The proof is a simple application of integration by parts. Consider

$$\int_\tau^{2\tau} \frac{2\tau - t}{\tau} \sin(tx)\, dt = \frac{\cos(\tau x)}{x} - \int_\tau^{2\tau} \frac{\cos(tx)}{\tau x}\, dt = \frac{\cos(\tau x)}{x} - \left(\frac{\sin(2\tau x) - \sin(\tau x)}{\tau x^2} \right).$$

Clearly $\left| \frac{\sin(2\tau x) - \sin(\tau x)}{\tau x^2} \right| \leq \frac{2}{\tau x^2}$ and so we have the desired result. □

Theorem D.9 (Cramer's inequality [33]). *Define the nth Hermite polynomial $He_n(x) = (-1)^n e^{\frac{x^2}{2}} \cdot \frac{d^n}{dx^n} e^{-\frac{x^2}{2}}$. Further define the Hermite function*

$$\psi_n(x) = (2^n n! \sqrt{\pi})^{-1/2} e^{-\frac{x^2}{2}} 2^{n/2} He_n(\sqrt{2}x).$$

Then $\sup_{x \in \mathbb{R}} |\psi_n(x)| \leq \pi^{-1/4}$.

Corollary D.10. *Let $h(x) = e^{-\frac{x^2}{2}}$. Then $|h^{(n)}(x)| \leq e^{-\frac{x^2}{4}} \sqrt{n!}$ for all $x \in \mathbb{R}$.*

Lemma D.11. *We have*

$$\sum_{k=0}^\infty \frac{1}{2^k k!} \left(\sum_{\ell=0}^k \binom{k}{\ell} \sqrt{\ell!} \right)^2 < \infty.$$

Further, for any fixed positive integer s, we have

$$\sum_{k=0}^\infty \frac{1 \vee k^s}{2^k k!} \left(\sum_{\ell=0}^k \binom{k}{\ell} \sqrt{\ell!} \right)^2 < \infty.$$

Proof. Expanding the square gives us

$$\frac{1}{2^k k!} \left(\sum_{\ell=0}^k \binom{k}{\ell} \sqrt{\ell!} \right)^2 = \frac{1}{2^k k!} \sum_{\ell=0}^k \binom{k}{\ell}^2 \ell! + \frac{1}{2^k k!} \sum_{0 \leq \ell \neq j \leq k} \binom{k}{\ell} \binom{k}{j} \sqrt{\ell! j!}.$$

Examining the first term, consider that

$$\frac{1}{2^k k!} \sum_{\ell=0}^k \binom{k}{\ell}^2 \ell! = \frac{1}{2^k} \sum_{\ell=0}^k \binom{k}{\ell} \frac{1}{(k-\ell)!} = \frac{1}{2^k} \sum_{\ell=0}^k \binom{k}{\ell} \frac{1}{\ell!} = \frac{1}{2^k} L_k(-1)$$

where L_k is the kth Laguerre polynomial. By the generating function $\frac{1}{1-t} e^{-\frac{tx}{1-t}} = \sum_{n=0}^\infty t^n L_n(x)$, we have

$$\sum_{k=0}^\infty \frac{1}{2^k} \sum_{\ell=0}^k \binom{k}{\ell} \frac{1}{\ell!} = \sum_{k=0}^\infty \frac{1}{2^k} L_k(-1) = 2e.$$

Thus the first term is handled and it remains to bound the cross term. Consider that for any $x, y \in \mathbb{R}$, we have $xy \leq x^2 \vee y^2$. Therefore,

$$\frac{1}{2^k k!} \sum_{0 \leq \ell \neq j \leq k} \binom{k}{\ell}\binom{k}{j} \sqrt{\ell! j!} \leq \frac{k}{2^k k!} \sum_{\ell=0}^{k} \binom{k}{\ell}^2 \ell! = \frac{k}{2^k} L_k(-1) \leq \left(\frac{3}{4}\right)^k L_k(-1)$$

where we have used $k \leq \left(\frac{3}{2}\right)^k$ for all $k \geq 0$. Employing the generating function as before, we have

$$\sum_{k=0}^{\infty} \frac{1}{2^k k!} \sum_{0 \leq \ell \neq j \leq k} \binom{k}{\ell}\binom{k}{j} \sqrt{\ell! j!} \leq \sum_{k=0}^{\infty} \left(\frac{3}{4}\right)^k L_k(-1) = 4e^3.$$

Thus, we have $\sum_{k=0}^{\infty} \frac{1}{2^k k!} \left(\sum_{\ell=0}^{k} \binom{k}{\ell} \sqrt{\ell!} \right)^2 \leq 2e + 4e^3 < \infty$ as desired. The second claim follows by a similar argument. $\square$

Lemma D.12 ([59]). *For probability measures $P_1, ..., P_n, Q_1, ..., Q_n$, we have*

$$\chi^2 \left(\bigotimes_{i=1}^{n} Q_i \,\bigg\|\, \bigotimes_{i=1}^{n} P_i \right) = \left(\prod_{i=1}^{n} (1 + \chi^2(Q_i \,\|\, P_i)) \right) - 1.$$

Theorem D.13 (Bounded differences - Theorem 6.2 [6]). *Suppose $f : \mathcal{X}^n \to \mathbb{R}$ satisfies the bounded differences inequality for some nonnegative $d_1, ..., d_n$, that is*

$$\sup_{\substack{x_1, ..., x_n \in \mathcal{X}, \\ x_i' \in \mathcal{X}}} |f(x_1, ..., x_{i-1}, x_i, x_{i+1}..., x_n) - f(x_1, ..., x_{i-1}, x_i', x_{i+1}, ..., x_n)| \leq d_i$$

for all $1 \leq i \leq n$. Let $Z = f(X_1, ..., X_n)$ where $X_1, ..., X_n$ are independent $\mathcal{X}$-valued random variables. If $u \geq 0$, then

$$P\{|Z - E(Z)| > u\} \leq 2 \exp\left(-\frac{2u^2}{\sum_{i=1}^{n} d_i^2} \right).$$

Theorem D.14 (Bernstein's inequality - Theorem 2.8.4 [60]). *Let $Y_1, ..., Y_n$ be independent mean zero random variables such that $|Y_i| \leq 1$ for all i. If $u \geq 0$, then*

$$P\left\{ \left| \sum_{i=1}^{n} Y_i \right| \geq u \right\} \leq 2 \exp\left(-\frac{u^2/2}{\tau^2 + u/3} \right)$$

where $\tau^2 = \sum_{i=1}^{n} E(Y_i^2)$.

E   Location estimation in Huber's contamination model

Recall Huber's contamination model (2), which is the same, from the perspective of conclusions about in-probability minimax estimation rates, as the Bayesian version (3). In [13], it has been shown that if $\varepsilon = \frac{k}{n} < \frac{1}{5}$, then the sample median achieves the optimal rate

$$|\text{median}(X_1, ..., X_n) - \theta|^2 \lesssim \sigma^2 \left(\frac{1}{n} + \varepsilon^2 \right) \asymp \sigma^2 \left(\frac{1}{n} + \frac{k^2}{n^2} \right).$$

In Remark 1 of Section 1.3, we claimed that for $\frac{n}{5} \leq k < \frac{n}{2}$,

$$|\text{median}(X_1, ..., X_n) - \theta|^2 \lesssim \sigma^2 \log\left(\frac{1}{1 - 2\varepsilon} \right) \asymp \sigma^2 \log\left(\frac{en}{n - 2k} \right).$$

and that this is the optimal rate. In this section, we give proofs for these two assertions. For convenience, the upper bound will be proved in the context of the frequentist model (2) while the lower bound will be proved in the context of the Bayes model (3) using the modulus of continuity concept of [13, 12].

The following proposition states the error achieved by sample median for $\frac{n}{5} \leq k < \frac{n}{2}$. The proof is

exactly the same as the proof of [38] in this regime, which we reproduce here with notational changes for the reader's convenience.

Proposition E.1. *Consider data $X_1, ..., X_n$ from the model (2). Suppose $\frac{n}{5} \le k < \frac{n}{2}$. If $\delta \in (0, 1)$ and n sufficiently large depending only on δ, then*

$$\sup_{\substack{\theta \in \mathbb{R}, \\ \eta \in \mathbb{R}^n, \\ \sigma > 0}} P_{\theta, \eta, \sigma} \left\{ |\text{median}(X_1, ..., X_n) - \theta| > C\sigma \sqrt{\log\left(\frac{en}{n - 2k}\right)} \right\} \le \delta$$

where $C > 0$ is a universal constant.

Proof. Without loss of generality we can assume $\sigma = 1$ otherwise we can simply work with the normalized data $\{X_j/\sigma\}_{j=1}^n$ because $\text{median}(X_1/\sigma, ..., X_n/\sigma) = \frac{\text{median}(X_1,...,X_n)}{\sigma}$. Let $\hat{\theta} := \text{median}(X_1, ..., X_n)$ We split into two cases.

Case 1: Suppose $n - 2k < \sqrt{n}$. Since $|\mathcal{I}| > \frac{n}{2}$, it follows that

$$P_{\theta, \eta, \gamma} \left\{ |\hat{\theta} - \theta| > \sqrt{4 \log(en)} \right\} \le P_{\theta, \eta, \gamma} \left\{ \max_{j \in \mathcal{I}} |X_j - \theta| > \sqrt{4 \log(en)} \right\} \le 2n \exp\left(-2 \log(en)\right) \le \frac{1}{n} < \delta$$

for $n \ge \frac{1}{\delta}$. Since $\log\left(\frac{en}{n-2k}\right) \asymp \log(en)$ because $n - 2k \le \sqrt{n}$, we have the claimed result.

Case 2: Suppose $n - 2k \ge \sqrt{n}$. Define the interval

$$E := \left[\theta - \sqrt{2 \log\left(\frac{4|\mathcal{I}|}{\frac{n}{2} - |\mathcal{O}|}\right)}, \theta + \sqrt{2 \log\left(\frac{4|\mathcal{I}|}{\frac{n}{2} - |\mathcal{O}|}\right)} \right].$$

Consider

$$P_{\theta, \eta, \sigma} \left\{ \hat{\theta} \notin E \right\} \le P_{\theta, \eta, \sigma} \left\{ \sum_{j=1}^n \mathbb{1}_{\{X_j \notin E\}} > \frac{n}{2} \right\}$$

$$\le P_{\theta, \eta, \sigma} \left\{ \sum_{j \in \mathcal{I}} \mathbb{1}_{\{X_j \notin E\}} > \frac{n}{2} - |\mathcal{O}| \right\}$$

$$\le P_{\theta, \eta, \sigma} \left\{ \sum_{j \in \mathcal{I}} \mathbb{1}_{\{X_j \notin E\}} - p > \frac{n}{2} - |\mathcal{O}| - |\mathcal{I}|p \right\}$$

where $p = P \left\{ |Z| > \sqrt{2 \log\left(\frac{4|\mathcal{I}|}{\frac{n}{2} - |\mathcal{O}|}\right)} \right\}$ with $Z \sim N(0, 1)$. By Bernstein's inequality (Theorem D.14), we have for a universal constant $c > 0$

$$P_{\theta, \eta, \sigma} \left\{ \sum_{j \in \mathcal{I}} \mathbb{1}_{\{X_j \notin E\}} - p > \frac{n}{2} - |\mathcal{O}| - |\mathcal{I}|p \right\} \le \exp\left(-c \min\left(\frac{\left(\frac{n}{2} - |\mathcal{O}| - |\mathcal{I}|p\right)^2}{|\mathcal{I}|p(1-p)}, \frac{n}{2} - |\mathcal{O}| - |\mathcal{I}|p \right) \right).$$

Consider that $p \le 2 \exp\left(-\frac{1}{2} \cdot 2 \log\left(\frac{4|\mathcal{I}|}{\frac{n}{2} - |\mathcal{O}|}\right) \right) = \frac{\frac{n}{2} - |\mathcal{O}|}{2|\mathcal{I}|}$. Therefore,

$$P_{\theta, \eta, \sigma} \left\{ \hat{\theta} \notin E \right\} \le \exp\left(-\frac{c}{2} \cdot \left(\frac{n}{2} - |\mathcal{O}|\right) \right) \le \exp\left(-\frac{c(n - 2k)}{4} \right) \le \exp\left(-\frac{c\sqrt{n}}{4} \right) \le \delta$$

for n sufficiently large depending on δ. Consider that $\log\left(\frac{4|\mathcal{I}|}{\frac{n}{2} - |\mathcal{O}|}\right) \lesssim \log\left(\frac{en}{n-2k}\right)$, and so the proof is complete. $\square$

For the discussion of the lower bound, let us consider the Bayesian model (3). The lower bound is proved through the use of a modulus of continuity [12, 13] which specializes in the context of our problem to

$$\omega(\varepsilon) = \sup \left\{ \frac{|\theta_1 - \theta_2|}{\sigma} : d_{\text{TV}}(N(\theta_1, \sigma^2), N(\theta_2, \sigma^2)) \le \frac{\varepsilon}{1 - \varepsilon} \text{ and } \theta_1, \theta_2 \in \mathbb{R}, \sigma > 0 \right\}.$$

Theorem E.2 (Theorem 5.1 [13]). *If $\varepsilon \in [0,1]$, then*

$$\inf_{\hat{\theta}} \sup_{\substack{\theta \in \mathbb{R}, \\ \sigma > 0, \\ Q}} P_{\varepsilon,\theta,Q,\sigma} \left\{ \frac{|\hat{\theta} - \theta|}{\sigma} \geq \frac{1}{\sqrt{n}} \vee \omega(\varepsilon) \right\} \geq c$$

for some universal constant $c > 0$.

A minimax lower bound thus follows by bounding the modulus of continuity from below.

Proposition E.3. *If $\varepsilon \in [0,1]$, then*

$$\omega(\varepsilon) \geq \sqrt{2 \log \left(\frac{(1-\varepsilon)/2}{1 - 2\varepsilon} \right)}.$$

Proof. From a well-known relationship between total variation and Kullback-Leibler divergence (e.g. [59]), we have that $d_{\mathrm{TV}}(N(\theta_1, \sigma^2), N(\theta_2, \sigma^2)) \leq 1 - \frac{1}{2} \exp\left(- \mathrm{KL}\left(N(\theta_1, \sigma^2) \,\|\, N(\theta_2, \sigma^2) \right) \right) = 1 - \frac{1}{2} \exp\left(-\frac{|\theta_1 - \theta_2|^2}{2\sigma^2} \right)$ where $\mathrm{KL}(\cdot \,\|\, \cdot)$ denotes the Kullback-Leibler divergence.

$$\begin{aligned}
\omega(\varepsilon) &\geq \sup \left\{ \frac{|\theta_1 - \theta_2|}{\sigma} : 1 - \frac{1}{2} \exp\left(-\frac{|\theta_1 - \theta_2|^2}{2\sigma^2} \right) \leq \frac{\varepsilon}{1-\varepsilon} \text{ and } \theta_1, \theta_2 \in \mathbb{R}, \sigma > 0 \right\} \\
&= \sup \left\{ \frac{|\theta_1 - \theta_2|}{\sigma} : \frac{2(1 - 2\varepsilon)}{1 - \varepsilon} \leq \exp\left(-\frac{|\theta_1 - \theta_2|^2}{2\sigma^2} \right) \text{ and } \theta_1, \theta_2 \in \mathbb{R}, \sigma > 0 \right\} \\
&= \sqrt{2 \log \left(\frac{(1-\varepsilon)/2}{1 - 2\varepsilon} \right)}.
\end{aligned}$$

The proof is complete. $\qquad \square$