# Analysis of the finite sample properties of the Power Enhancement technique

**Stan Koobs**

A thesis presented for the degree of
*Bachelor of Science*


University of Groningen
Faculty of Economics and Business

August 14, 2020

Bachelor's Thesis Econometrics and Operations Research

Supervisor: N.W. Koning, MSc
Second assessor: prof. dr. R.J.M. Alessie

university of groningen / faculty of economics and business

# Analysis of the finite sample properties of the Power Enhancement technique

August 14, 2020

**Author**: Stan Koobs
**Supervisor**: Nick Koning, MSc

## Abstract

In this paper, we investigate the finite sample properties of the recently introduced power enhancement technique. As has been shown in the literature, this technique has outstandingly strong asymptotic results. Yet, it remains unclear how these results carry over to samples of a practical size. Before diving into this problem, this paper will first give an overview of the background, in which we elaborate on the concepts of high-dimensional testing, sparsity, and the power enhancement technique itself. Subsequently, we perform a theoretical analysis. In this analysis, we reveal the finite sample properties of the technique in a basic setting. Afterwards, we numerically verify and extend these results by using Monte Carlo simulations. We discovered some substantial size distortion and the technique is highly sensitive to the choice of the threshold. Therefore, the practitioner should be cautious when applying this method.

# 1 Introduction

The emergence of big data brings a lot of opportunities and challenges to the modern world. On the one hand, it gives fields like machine learning the possibility to discover small patterns that were not detectable in small-scale data. On the other hand, this tremendous amount of data also brings a set of unique statistical problems. Consider the case of biology and specifically the field of genetics. Here, millions of genes are measured for a single individual. Consequently, the number of features measured largely surmounts the number of observations. However, most conventional statistical techniques are developed for the setting of a small number of parameters and a lot of observations. A consequence is that the performance of the test is often deteriorated when the amount of parameters becomes relatively large.

To examine the performance of such a test, two properties are crucial. These are the size and the power of a test which are defined by:

$$\begin{aligned} \text{Size} &= \mathbb{P}(\text{reject } H_0 \mid H_0 \text{ is true}), \\ \text{Power} &= \mathbb{P}(\text{reject } H_0 \mid H_1 \text{ is true}). \end{aligned}$$

In an ideal world, the size is as small as possible and that the power is as high as possible. However, there is a trade-off here: when the size decreases, the power decreases generally too. To handle this, the Neyman-Pearson approach is often used: fix the size and maximize the power.

In this study, we are interested in the size and power when testing a hypothesis on a high-dimensional parameter:

$$H_0 : \boldsymbol{\theta} = \mathbf{0} \quad \text{against} \quad H_1 : \boldsymbol{\theta} \neq \mathbf{0}.$$

Here, the dimension of $\boldsymbol{\theta}$, denoted by $p$, may exceed the number of observations $n$. The conventional statistical techniques often lose a lot of power when testing this hypothesis when $p$ is large. The intuition behind this is that as we keep increasing $p$, the parameter space of $H_1$ becomes larger and larger but the parameter space of $H_0$ stays at the origin. Since the space of alternatives keeps increasing, most tests have trouble detecting these alternatives which yields a low probability of rejection.

In this paper, a method to boost the power in this setting will be investigated. We will study the power enhancement technique suggested by Fan, Liao, and Yao (2015). The idea of this technique is to increase the power of an existing test while the size stays the same or minimally increases. The work of Fan et al. (2015) and Kock and Preinerstorfer (2019) especially promises very strong asymptotic results for this technique. To what extent this result carries over to a finite sample, is still unknown. In this paper, we investigate the performance of this test in samples of a practical size.

The technique behind this method is fairly simple. First, an initial test is taken which has a correct asymptotic size. An example of a test that can be used here is the Wald test. The test statistic that belongs to this initial test is denoted by $J_1$. The power enhancement test statistic is then denoted by $J$ and we reject the hypothesis if $J > c$ where $c$ is the critical value. It is defined by:

$$J = J_0 + J_1,$$

where $J_0$, also known as the power enhancement component, is another test statistic that satisfies three special properties.

The first one of these is that it should be non-negative. Because the test has a right-tailed rejection region, adding $J_0$ to the initial test can only increase the probability of rejection. Therefore, this property guarantees that the power of the test is at least as large as the power of the initial test. The second property states that the probability that $J_0 = 0$ should converge to 1 under the null hypothesis. Hence, this property ensures that the test asymptotically has the same size as the initial test. Notice that this property is essential for the asymptotic results to hold since this property assures that the asymptotic distribution of $J$ under the null hypothesis is entirely determined by $J_1$. So, the first two properties make sure that the test never performs worse than the initial test in terms of power and that the size is the same asymptotically. Finally, the third property gives the test more power in specific regions of the alternative $H_1$. It states that $J_0$ should diverge in probability under some specific regions of alternatives $H_1$ where the initial test is inconsistent. As shown in the work of Fan et al. (2015), there exist $J_0$ components which satisfy all these conditions.

Most existing tests used to test the hypothesis $H_0 : \boldsymbol{\theta} = \mathbf{0}$ have a quadratic form like the Wald statistic. Moreover, Fan (1996) showed that in this $p > n$ setting, the Wald test even suffers from low power in a simple case of testing on the mean of a normal distribution. These tests are based on the $\ell_2$-norm of a vector, also called the Euclidean length of a vector. Consequently, these tests then have good power against alternatives that have a high $\ell_2$-norm. In the paper of Fan et al. (2015), $J_0$ is chosen such that it enhances the power in the regions in which only a few components of $\boldsymbol{\theta}$ highly violate the null hypothesis. These are the so-called 'sparse' alternatives. For these sparse alternatives, the $\ell_2$-norm stays relatively low compared to the length of the vector. Therefore, most existing quadratic tests have low power against these alternatives. Now, the power enhancement technique can be applied to boost the power against these sparse alternatives.

This boost in power while keeping the size distortion small especially works well asymptotically. In the work of Kock and Preinerstorfer (2019), these asymptotic properties have been analyzed. In their paper, they showed that in the setting where the dimension of the parameter vector remains fixed, there exist tests that cannot be further improved by the power enhancement technique. However, this changes drastically when the dimension of the parameter vector is also allowed to increase. In that case, they showed that for all sufficiently slowly increasing growth rates of the dimension of the parameter vector *every* test with an asymptotic size less than one will obtain more power. Yet, it remains unclear how these results carry over to finite samples. Therefore, this study will focus on analyzing these finite sample properties.

One compelling example of an application of this technique can be found in financial econometrics. A model that is often used to predict the excess returns of stocks on a market is the factor pricing model. In this model, we put a statistical structure on the excess return and try to explain it using factors like the return of industry portfolios and the book-to-market ratio of a company. This model was first introduced by Fama and French (1992). In this model, we let $r_{it}$ denote the excess return of stock $i$ at time $t$, $\boldsymbol{f_t}$ denotes the $K$-dimensional vectors of observable factors at time $t$, and $\boldsymbol{b_i}$ denotes the $K$-dimensional vector of factor loadings assigned to stock $i$. The model is then given by:

$$r_{it} = \theta_i + \boldsymbol{b_i^T f_t} + e_{it}, \quad i = 1, \ldots, p, \quad t = 1, \ldots, n,$$

where $\theta_i$ is the intercept belonging to stock $i$ (which is also known as the 'alpha' of stock $i$ in finance literature) and $e_{it}$ is the idiosyncratic error of stock $i$ at time $t$. So, in this model, we have the stocks of $p$ firms and on each of these stocks, we have $n$ observations. The

important implication from factor pricing theory is that the intercept $\theta_i$ should be zero for all stocks. If $\theta_i$ would be positive for some stock $i$, the stock is overpriced according to the model. This implication of all intercepts being zero is also called 'mean-variance efficiency'. Therefore, to check if this mean-variance efficiency holds in a certain market we would test the following hypothesis:

$$H_0 : \boldsymbol{\theta} = \mathbf{0} \quad \text{against} \quad H_1 : \boldsymbol{\theta} \neq \mathbf{0},$$

where $\boldsymbol{\theta}$ is the $p$-dimensional vector of intercepts for all stocks. This is exactly the hypothesis on which this research is focused and for which the power enhancement method might be convenient. This study can give further insights into the usefulness of the power enhancement method when the number of firms is relatively large in this model. It turns out that this sparse structure can also be found in this example. In an empirical study Fan et al. (2015) found evidence that only a few stocks of the S&P 500 are significantly mispriced according to their factor pricing model instead of the whole market being structurally mispriced. Next to this example, sparsity is also seen in a lot of other fields completely different from finance like physics. Since this is such a relevant concept in practice, we will start this paper with a detailed discussion of this notion.

So, the power enhancement technique seems a promising technique with useful applications. According to the asymptotic results, you acquire the power almost for 'free'. However, of course, everything has a prize and we are interested in what this prize is for 'realistic settings'. In this paper, we will investigate the technique for these more or less realistic sample sizes. We will do so by using a theoretical analysis and by Monte Carlo simulations.

The remainder of this paper is structured as follows. Section 2 contains the background of this study. This section summarizes the recent literature and some essential concepts like sparsity and high-dimensional testing. Afterwards, Section 3 discusses the technique in full detail. When the background and technique have clearly been explained, Section 4 formulates the problem addressed in this study. This problem will be tackled in Section 5 by a theoretical analysis. Subsequently, Section 6 verifies and extends these results by using Monte Carlo simulations. Finally, Section 7 draws a conclusion.

# 2 Background

In this section, we first clarify some concepts which are essential for our research. We start off by elaborating on the notion of sparsity. Here we will discuss what practitioners often desire of a sparsity measure and some existing measures which are being used. Subsequently, we consider the field of high-dimensional testing and why sparsity is often relevant here in practice. Thereafter, we consider the drawbacks of applying conventional tests in this sparse setting. Finally, some existing tests designed to address these drawbacks are presented.

## 2.1 Sparsity

Surprisingly, while sparsity is an increasingly commonly used concept in the econometrics and statistics literature, there does not seem to be consensus on what sparsity actually is. Practitioners often used sparsity to demonstrate how many components are zero compared to how many components are non-zero. Sometimes it is also used more generally: a vector is called 'sparse' when the Euclidean length of a vector is determined by only a small number

of elements. This issue of a lack of consensus has also been addressed by Hurley and Rickard (2008). In their study, they compared several measures of sparsity and they did this based on a set of axioms. These six axioms can be seen as rules of which the practical user thinks that should be fulfilled by a good sparsity measure. They proposed the following axioms:

- *Robin Hood:* decreasing the value of large signals and increasing the value of the small signals reduces sparsity.
- *Scaling:* if a vector is multiplied by a constant, the sparsity should not change. This is equivalent to saying that the sparsity measure should be a function which is homogeneous of degree zero. In some sense, we can speak of sparse 'lines' through the origin.
- *Rising tide:* adding a constant to each component reduces sparsity. This axiom assumes that the vector contains at least two components with different values. Therefore, adding a constant reduces the relative difference.
- *Cloning:* duplicating data preserves sparsity.
- *Bill Gates:* one component becoming large increases sparsity.
- *Babies:* adding zeros to a vector increases sparsity.

As also shown in the work of Hurley and Rickard (2008), these axioms are actually slight modifications of axioms introduced in an economics paper by Dalton (1920). In this paper, these axioms were used to measure the inequality of income in a society.

In the following subsections, we will go through some measures of sparsity and discuss why or why not these would be useful sparsity measures in practice.

### 2.1.1 The $\ell_0$-norm

Some traditional measures of sparsity use the $\ell_0$-norm. This $\ell_0$-norm gives the number of non-zero elements a vector. Now, let $\boldsymbol{\theta}$ be a vector of length $p$. One example of a sparsity measure which uses this norm is the following:

$$\frac{p - ||\boldsymbol{\theta}||_0}{p},$$

which is basically the number of zero elements divided by the total number of elements. One problem with this measure is that it treats large numbers the same as small numbers which differ only slightly from zero. For example, according to this measure, the vector $x = (0.01, 0.01, 0)$ is just as sparse as the vector $y = (5, 0.01, 0)$. In practice, this is generally undesirable. In most settings, there is some noise and therefore small effects are just as irrelevant as no effects. This is also the reason why this measure violates the *Robin Hood* and the *Bill Gates* property. Moreover, it can easily be seen that it also violates the *Rising tide* property.

Nevertheless, consider the case in which we are only comparing vectors consisting of zeros and non-zero elements of comparable absolute magnitude. Then these three axioms which are violated are actually irrelevant. Therefore, this measure could be interesting to use in that setting.

### 2.1.2 Hoyer index

Although the $\ell_0$-norm is not always that useful, there may be some valuable information in higher norms. Let us consider the $\ell_1$-norm which is the sum of absolute values of a vector and the $\ell_2$-norm which is the Euclidean length of a vector. Now, for simplicity, consider the two-dimensional case. In Figure 2.1, a level plot of these two norms can be found. In this plot, all two-dimensional vectors $\boldsymbol{\theta}$ which satisfy $||\boldsymbol{\theta}||_1 = 1$ or $||\boldsymbol{\theta}||_2 = 1$ can be seen.

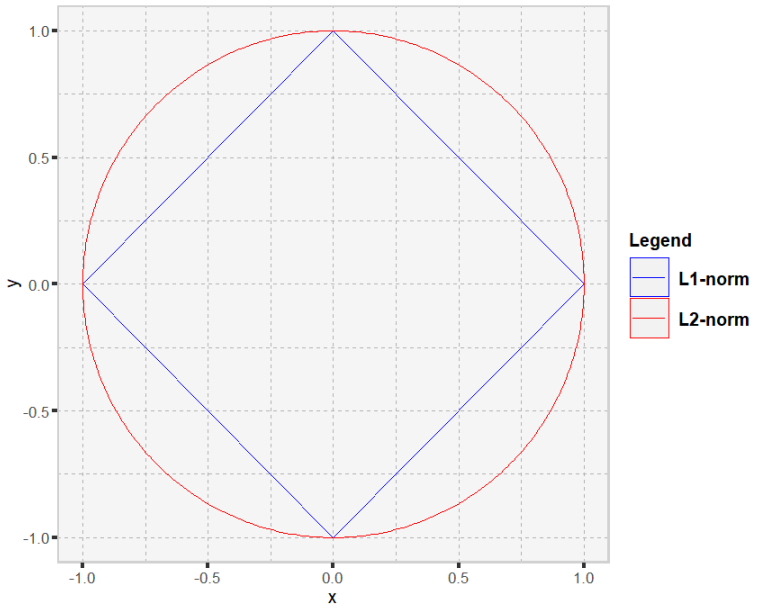


Figure 2.1: Level curves of the $\ell_1$-norm and $\ell_2$-norm of value 1

According to the axioms already stated, a vector with one element equal to one and the other elements equal to zero, is seen as a very sparse vector. On the contrary, a vector with all elements equal to each other is regarded as a very dense vector. Now, note that in the above figure, the two norms are only equal to each other when $x = 0$ or $y = 0$. These are exactly the vectors which are regarded as very sparse. Furthermore, the distance between these two norms becomes larger when we move to the more dense vectors. The distance between them is at its maximum on when $y = x$ or when $y = -x$ which is exactly the case where vectors are regarded as very dense. So, it seems like this ratio of the $\ell_1$-norm and the $\ell_2$-norm can tell us something about the sparsity of a vector. This is also the idea which the Hoyer index uses, which is also one of the measures discussed by Hurley and Rickard (2008). This measure is defined as:

$$H(\boldsymbol{\theta}) = \frac{\sqrt{p} - \frac{||\boldsymbol{\theta}||_1}{||\boldsymbol{\theta}||_2}}{\sqrt{p} - 1}.$$

When considering vectors of dimension $p$, multiples of the canonical basis vectors $\boldsymbol{e}_i$ $(1 \leq i \leq p)$ where the $i$th coordinate is equal to 1 and all other coordinates are equal to zero, are regarded as the 'most' sparse vectors. Now, let $\alpha$ be a positive constant. Plugging such

a vector into the index gives:

$$\begin{aligned} H(\alpha \boldsymbol{e}_i) &= \frac{\sqrt{p} - \frac{||\alpha \boldsymbol{e}_i||_1}{||\alpha \boldsymbol{e}_i||_2}}{\sqrt{p} - 1} \\ &= \frac{\sqrt{p} - \frac{\alpha}{\alpha}}{\sqrt{p} - 1} \\ &= 1. \end{aligned}$$

As it turns out, $H(\boldsymbol{\theta}) = 1$ can also be rewritten to $||\boldsymbol{\theta}||_1 = ||\boldsymbol{\theta}||_2$ which means that this index is equal to 1 for all vectors of which the $\ell_1$-norm equals the $\ell_2$-norm. Now consider a dense vector which satisfies $\theta_1 = \ldots = \theta_p = \theta$. Plugging this in yields:

$$\begin{aligned} H(\boldsymbol{\theta}) &= \frac{\sqrt{p} - \frac{||\boldsymbol{\theta}||_1}{||\boldsymbol{\theta}||_2}}{\sqrt{p} - 1} \\ &= \frac{\sqrt{p} - \frac{p|\theta|}{\sqrt{p|\theta|^2}}}{\sqrt{p} - 1} \\ &= \frac{\sqrt{p} - \sqrt{p}}{\sqrt{p} - 1} \\ &= 0. \end{aligned}$$

So, this Hoyer index gives a measure between 0 and 1 which shows how sparse a vector is. Now it might be interesting to see how this measure behaves in the two-dimensional case. In Figure 2.2, a heat map of the Hoyer index is shown.

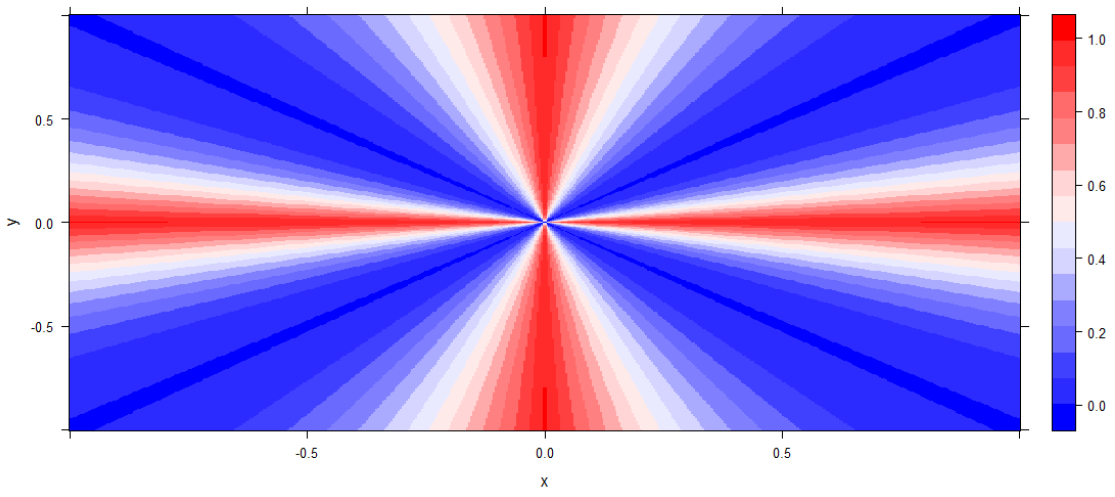


Figure 2.2: Heat map of the Hoyer index

This ratio of the $\ell_1$-norm and the $\ell_2$-norm as can be seen in the Hoyer index is also investigated by Yin, Esser, and Xin (2014), who dive a lot deeper into this concept. When looking at our axioms, this Hoyer index is an improvement compared to a measure which only uses the $\ell_0$ norm. As will be explained, it satisfies all axioms except the *cloning* and

*babies* axiom. The latter of the two is only violated if we are looking at a multiple of a standard basic vector.

Why this first axiom does not hold can be easily seen when we consider the vectors $(1, 0)$ and $(1, 0, 1, 0)$. According to the *cloning* property, these vectors are equally sparse. However, according to the Hoyer index, the first vector has a sparsity of 1 and the second vector has a sparsity of $2 - \sqrt{2} \approx 0.59$. This axiom can be relevant in some settings of our study.

The second axiom is even more relevant in our study. Consider the case where we look at a vector $\boldsymbol{\theta} = (1, 0, 0)$. When we start increasing the dimension $p$ by adding zeros to it, this vector becomes sparser and sparser. Yet, when we compare these vectors using the Hoyer index, they all just have index 1 and they are equally sparse. This is a relevant case if we want to investigate to what extent the power enhancement technique is able to pick up this violation in the first component. Notice, that this *babies* axiom is actually satisfied for all other vectors except the scalar multiples of the standard basis vector.

Nevertheless, these two violated properties do only make sense if we want to compare vectors of different lengths. So, the Hoyer index is a very useful measure of sparsity, except when we are comparing vectors with a different number of elements. In the next section, we consider an index that is compatible with that setting.

### 2.1.3 Gini index

In their study, Hurley and Rickard (2008) suggested one sparsity measure which satisfies all six axioms. This sparsity measure is called the Gini index. The name of this index comes from the Gini coefficient which is often used to measure income inequality. Let us denote $|\theta|_{(i)}$ for the $i$th smallest element in absolute value of a vector $\boldsymbol{\theta}$ $(1 \leq i \leq p)$. So, after taking the absolute value and then sorting this vector, we get the vector $(|\theta|_{(1)}, |\theta|_{(2)}, \ldots, |\theta|_{(p)})$. Then the Gini index is defined as:

$$G(\boldsymbol{\theta}) = 1 - 2 \sum_{i=1}^{p} \frac{|\theta|_{(i)}}{||\boldsymbol{\theta}||_1} \left( \frac{p - i + \frac{1}{2}}{p} \right).$$

Here, we again see the relation with the Gini coefficient from economics. The only difference with the economic Gini coefficient is that the economic one uses $\frac{p+1}{p}$ for the 1 in front and that it adds 1 instead of $\frac{1}{2}$ in the weighting factor.

The idea here is that we again have a weighted sum of components just like the Hoyer index but unlike the $\ell_0$-norm. In the Gini index, the weights are determined by the order of the components. Here the small components get more weight than the large ones. This is to make sure that a change in a large coefficient does not overwhelm the effect of a change in a small component. Furthermore, this index also uses some normalization. Dividing by the $\ell_1$-norm is done to make sure that a vector is not deemed more or less sparse because it has louder or quieter coefficients. It all depends on the relative values of the components compared to the total value. Notice that this cannot be achieved by the $\ell_2$-norm since this norm is too small and therefore it could lead to a negative sparsity. The weight is also divided by $p$ to make sure there is also some normalization regarding the number of elements of the vector since we also want to be able to compare vectors with a different number of elements. The outcome of this measure also lies between 0 and 1.

## 2.2 High-dimensional testing

Before we can connect sparsity to the power enhancement technique, we first need to clarify another essential concept. The power enhancement technique is based upon the field of high-dimensional testing. This field concerns the testing of hypotheses on high-dimensional vectors where the number of parameters is higher than the number of observations. This brings some challenges when applying classical statistical techniques and in this section, we will elaborate on these challenges.

Most classic tests to examine our hypothesis are based on a quadratic form. Because of this, these tests possess a lot of power against alternatives that have a relatively high $\ell_2-$norm. For a simple null hypothesis, the quadratic test statistic has a form which looks like:

$$W = \widehat{\boldsymbol{\theta}}^{\boldsymbol{T}} \boldsymbol{V} \widehat{\boldsymbol{\theta}},$$

where $\widehat{\boldsymbol{\theta}}$ is a consistent estimator of $\boldsymbol{\theta}$ and $\boldsymbol{V}$ is a positive definite weight matrix. One example of such a test is the Wald test. In this test, the inverse of the asymptotic covariance matrix of $\widehat{\boldsymbol{\theta}}$ is used for $\boldsymbol{V}$. In practice, this matrix is usually unknown and must be estimated. However, this is problematic in a high-dimensional setting as the sample analogue of the covariance matrix is singular and therefore, not invertible.

Another quadratic statistic is given by the Hotelling $T^2$-test statistic. This test is often used when testing whether the mean vector of a population is zero or testing whether the mean vectors of two populations are equal. Applying this test in a high-dimensional setting also causes problems since the test statistic does not exist when $p > n$. To counter this problem, multiple other test statistics are proposed by Dempster (1958), Bai and Saranada (1996), Srivastava and Du (2008), and Chakraborty and Chaudhuri (2017).

This problem of having more parameters than observations also arises in regression models where the conventional F-test is no longer applicable, replacements are considered by Zhong and Chen (2011) and Steinberger (2016).

As also shown by Fan (1996), most quadratic tests suffer from low power when $p > n$. This low power happens if the $\ell_2$-norm does not grow fast enough with $p$. Especially the sparse alternatives have a relatively low $\ell_2$-norm since there are only a few components which highly violate the null compared to the total number of elements. Before we dive deeper into this, we will first motivate why these sparse alternatives are so relevant in a high-dimensional setting.

## 2.3 Sparsity in practice

There are many high-dimensional situations of practical interest in which sparsity is highly relevant. One example of this can be found in panel data models. Currently, the power enhancement technique is already being used for multiple purposes here. Let $y_{it}$ denote the value of the dependent variable for individual $i$ at time $t$ and let $\boldsymbol{x_{it}}$ denote the vector with observed characteristics of this individual $i$ at time $t$. The model then has the following form:

$$y_{it} = \alpha + \boldsymbol{x}_{it}^T \boldsymbol{\beta} + c_i + u_{it}, \quad i = 1, \ldots, n, \quad t = 1, \ldots, T,$$

where we want to estimate $\alpha$ and $\boldsymbol{\beta}$. Next to that, $c_i$ is the individual effect and $u_{it}$ is the idiosyncratic error. In such a model, we often assume that $u_{it} \overset{i.i.d.}{\sim} (0, \sigma_u^2)$ where $\sigma_u^2$ is the variance which is constant among all errors. This does not always hold in practice and one

of the things which we might want to validate is the independence of individuals. This is also called the cross-sectional independence. Let us denote $\rho_{ij}$ for the correlation between $u_{it}$ and $u_{jt}$, which is assumed to be time-invariant. The following hypothesis can now be used to test cross-sectional independence:

$$H_0 : \rho_{ij} = 0 \quad \text{against} \quad H_1 : \rho_{ij} \neq 0, \qquad \text{for all } i \neq j.$$

If we would then define $\boldsymbol{\theta} = (\rho_{12}, \ldots, \rho_{n-1,n})$, which contains all elements above the diagonal from the covariance matrix of $\boldsymbol{u_t}$ which is the vector of idiosyncratic shocks at time $t$, we could test $H_0 : \boldsymbol{\theta} = \mathbf{0}$ to test this. It is often the case here that $\boldsymbol{\theta}$ is sparse, as also stated by Fan et al. (2015). There are just a few individuals who are correlated but there is no structural correlation among all individuals. Therefore, the power enhancement technique can be very convenient in this situation to raise the power. This has also been done by Fan et al. (2015) and a refinement of this can be found in Juodis and Reese (2018) who used a slightly different form for the power enhancement component $J_0$. Next to that, Su, Zhang, and Wei (2016) suggest using the power enhancement technique for testing the strict exogeneity assumption which is the assumption that states that the expectation of idiosyncratic error $u_{it}$ conditional on the observed characteristics and the individual effects, is zero. If this assumption does not hold, the FE and FD estimators are inconsistent so it is crucial to have a powerful test here for model validation.

Another challenging problem in econometrics is testing the independence of several high-dimensional vectors. In the work of Chen and Y. (2019), a new test statistic is proposed which also makes use of the power enhancement method. The power enhancement technique can be helpful in this situation when there are only a few components which are dependent.

Furthermore, sparsity is also highly relevant in biostatistics. In the field of genetic pathway analysis, Liu, Sun, Alexander, Kooperberg, and He (2019) used the power enhancement method to detect pathways that have sparse signals. Moreover, Xu, Lin, Wei, and Pan (2017) used the technique for two-sample testing of two high-dimensional means in genomics and genetics. Sparsity is so pertinent since each phenotype is likely to be influenced by a small number of genes, rather than all the genes. Most existing tests have very low power in this setting since there are millions of genes measured and the number of individuals is a lot lower. This is also motivated by (Wang, Yang, and Deng, 2015).

This sparse structure can also be found in other fields like physics. It can be seen in applications like ultrasonic flaw detection in highly scattering materials (Zhang, Zhang, and Wang, 2000) or the detection of hydrocarbons in materials (Castagna, Sun, and Siegfried, 2003). Another interesting application is medical imaging, where MRI scans are often used to detect breast cancer. These scans are used to visualize microcalcifications which might be a sign of breast cancer. These signals are very uncommon in the breast tissue (James, Clymer, and Schmalbrock, 2001), which again shows the sparse structure.

## 2.4 Power of traditional tests against sparse alternatives

As we saw in the previous subsection, sparsity is highly relevant in a lot of high-dimensional settings of practical interest. Nevertheless, conventional tests have very lower power against these alternatives. In this subsection, we will elaborate on why this is the case.

We say that a test is "consistent" against a specific alternative if its power converges to 1 as the number of observations increases. But in this high-dimensional setting, $p$ also

keeps increasing as the number of observations increases. We will use $p(n)$ to denote the value of $p$ as a function of $n$. The growth of $p$ is often not a problem if the alternative is dense. In that case, all components have a significant contribution to the $\ell_2$-norm such that it grows fast when $p$ increases. For example, in a normal location model, a condition for a simple $\chi^2$ test to be consistent against an alternative $\boldsymbol{\theta}$, is that $\frac{n}{\sqrt{p(n)}}||\boldsymbol{\theta}||_2^2$ diverges (Kock and Preinerstorfer, 2019). For dense alternatives of which the elements also grow fast enough, this condition generally holds.

Let us now consider an example with a sparse alternative. We define the $p(n)$-dimensional vector:

$$\boldsymbol{\theta}_n = (\log p(n), 0, \ldots, 0).$$

As we start increasing $n$, we are adding zeros to this vector and the first component becomes larger and larger. Let us consider an example in which this vector $\boldsymbol{\theta}$ is the mean of a multivariate normal distribution. By assuming that the covariance matrix is equal to the identity matrix, it follows that the Wald test statistic to test $H_0 : \boldsymbol{\theta} = \mathbf{0}$ has a non-central $\chi^2$ distribution under the alternative. Furthermore, we assume that $p$ grows exponentially with $n$ in the following way: $p(n) = 2^n$. Since we know the finite sample distribution of the test statistic, we can compute the power against this alternative. There are two counteracting effects on the power here as we start increasing $n$. The first one follows from the $\ell_2-$norm of $\boldsymbol{\theta}$. Due to the first component, this norm is increasing in $n$. The second effect follows from the critical value of this test. Since the null distribution is given by a $\chi^2$ distribution with $p(n)$ degrees of freedom, this critical value also increases with $n$. This is illustrated in Figure 2.3.

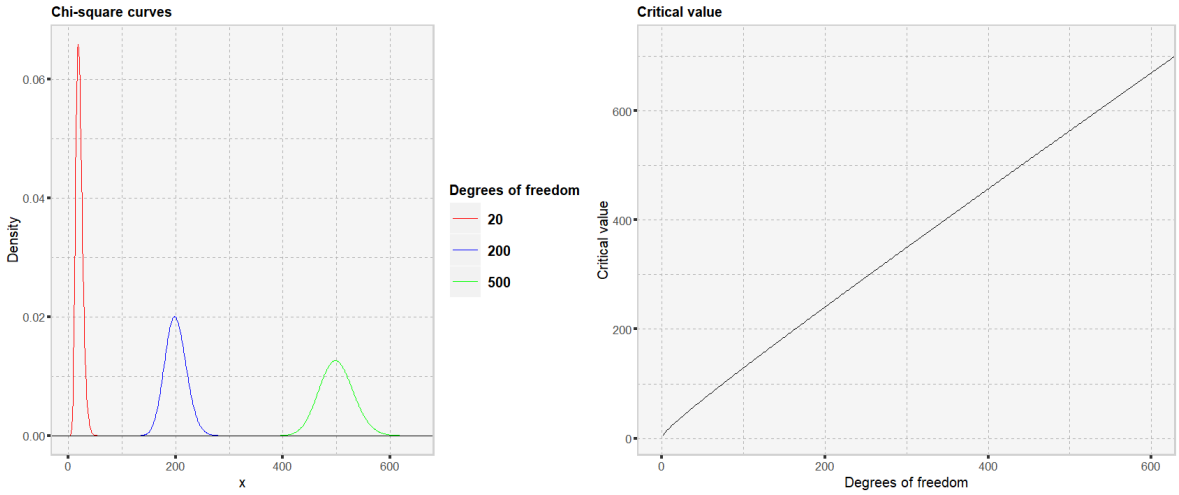


Figure 2.3: Critical value of the $\chi^2$ distribution

So, the power of this test depends on the interaction of the $\ell_2$-norm and the critical value. In Figure 2.4, the power is shown against the value of $p(n)$.

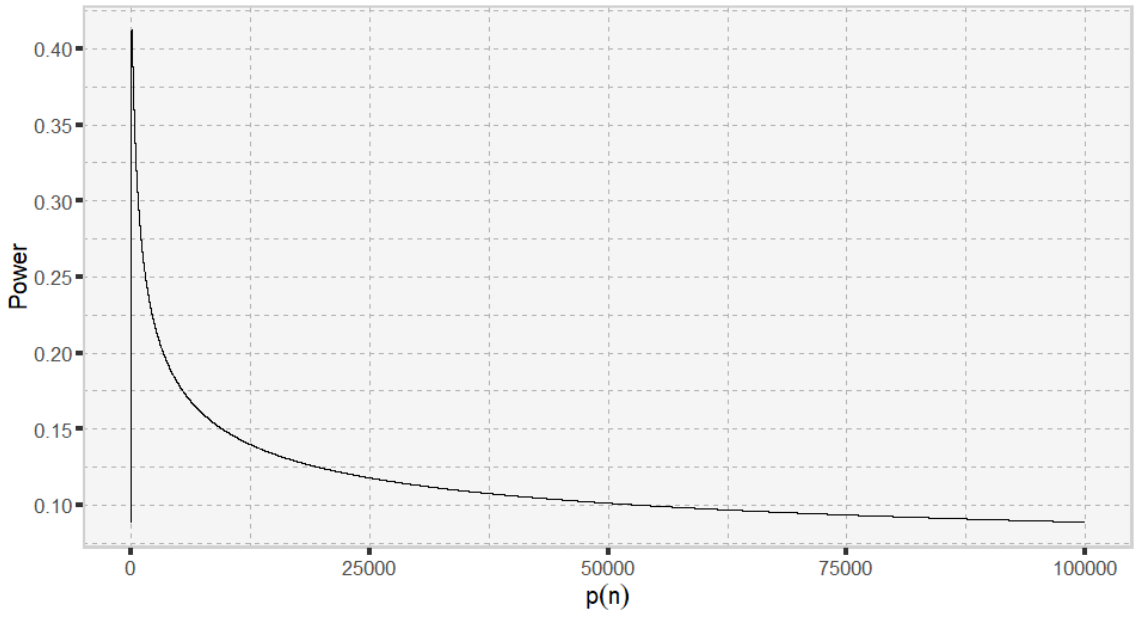


Figure 2.4: Power of the Wald test

We observe in this figure that the power increases first, this is due to the $\log(p(n))$ term which is a concave function which increases relatively fast for small $p(n)$. After some point, the critical value starts to dominate and the power starts to decrease. Eventually, the power converges to the significance level. This shows that the conventional test indeed has low power against this alternative.

Since we are considering a normal location model here, the inconsistency against this alternative also follows from the condition given by Kock and Preinerstorfer (2019). We find that:

$$\frac{n}{\sqrt{p(n)}}||\boldsymbol{\theta}||_2^2 = \frac{n}{2^{n/2}}(\log(2^n))^2 = \frac{n^3\log(2)^2}{2^{n/2}} \to 0,$$

where the last step follows from the fact that the numerator grows polynomially in $n$ and the denominator exponentially in $n$. This example illustrates the low power of a quadratic test against a sparse alternative.

In their paper, Fan et al. (2015) explained this phenomenon as the 'accumulation of estimation errors'. When we are testing against a sparse alternative, there are a lot of components that have a negligible effect on the $\ell_2$-norm. Nevertheless, these components need to be estimated and therefore, they are taken into account by the $\chi^2$ distribution as an extra degree of freedom. Consequently, it might happen that the critical value of the test dominates the sparse signals which gives the test low power against these alternatives.

## 2.5 Existing tests

The low power of the traditional tests against sparse alternatives and the practical relevance of sparsity demonstrates the demand for tests that are able to enhance the power against these alternatives. In the existing literature, there have been some other proposed testing techniques that satisfy this criterion. The main two types of tests which satisfy this are the extreme value test and the thresholding test. The idea of the extreme value test is based

on the maximum deviation from the null where the statistic is based on $\max_{j\leq p}|\frac{\widehat{\theta}_j}{w_j}|^\delta$ where $w_j$ is the weight and $\delta > 0$ is a positive constant. Examples of this test can be found in Chernozhukov, Chetverikov, and Kato (2013), Cai, W.Liu, and Y.Xia (2013) and Cai, Liu, and Xia (2014). Another test which is interesting to compare with is the test based on thresholding. In this test, the test statistic is:

$$R = \sqrt{n}\sum_{j=1}^{p}\left|\frac{\widehat{\theta}_j}{w_j}\right|^\delta \mathbb{I}\{|\widehat{\theta}_j| > t_p w_j\}$$

where $t_p$ is the threshold level and $\mathbb{I}$ is used as an indicator variable which is equal to 1 if the condition inside the brackets is satisfied. So this test statistic is the sum of $\left|\frac{\widehat{\theta}_j}{w_j}\right|^\delta$ terms for every $j \in \{1,\ldots,p\}$ of which $\frac{|\widehat{\theta}_j|}{w_j}$ exceeds the threshold $t_p$. The idea of this test is also used by Hansen (2005) and Zhong, Chen, and Xu (2013).

Another test that has high power against sparse alternatives is the test given by Koning (2019). This test statistic generalizes the Wald statistic and directs the power towards a conic parameter subspace chosen by the user. In the paper, this test statistic is illustrated on subspaces that consist of sparse or nearly-sparse vectors.

The existing tests also certainly have some disadvantages. As shown by Hansen (2003), the statistics of extreme value and thresholding tests are nonpivotal and require bootstrap to derive their null distribution. According to Fan et al. (2015), these tests have a slow rate of convergence which makes the asymptotic null distribution inaccurate in a finite sample. Furthermore, the disadvantage of bootstrap is that it can be very time-consuming. In all, this shows that the existing tests clearly have some drawbacks. Therefore, it is highly pertinent to further investigate the power enhancement technique and inspect how this technique performs in practice.

# 3 Layout of the technique

As shown in the previous section, it is highly relevant in practice to have a test that is able to boost the power against sparse alternatives. In this section, we will describe the power enhancement technique introduced by Fan et al. (2015) in detail. This provides us some necessary insights which we can use in our analysis later. We will first give an example of a statistic that can be used as power enhancement component. Subsequently, in Section 3.3, we will verify the use of the critical value of the initial test as the critical value of the power enhancement test. Finally, we will demonstrate what happens when we use this test against the sparse example from Section 2.4.

In this technique, we take an initial test that already has correct asymptotic size. This initial test statistic is denoted by $J_1$. The power enhancement test statistic is then defined by:

$$J = J_0 + J_1,$$

where $J_0$ is the power enhancement component. There are several options for the power enhancement component but all of these options have to satisfy the following properties:

(a) $J_0 \geq 0$ almost surely,

(b) $\mathbb{P}(J_0 = 0|H_0) \to 1$,

(c) $J_0$ diverges in probability under some specific regions of alternatives $H_1$.

We will now first give an example of a statistic that can be used as power enhancement component and after that, we will verify the above conditions.

For all asymptotic results, we assume that the dimension $p$ is a function of the number of observations $n$ here and that it is allowed to grow much faster than $n$. In the literature, it is often taken that $p$ grows polynomially or exponentially with $n$ (Fan and Tang, 2012).

## 3.1 The screening statistic

Let us use $v_j$ to denote the asymptotic variance of $\widehat{\theta}_j$ and $\widehat{v}_j$ to denote its estimator. An example of $J_0$ proposed by Fan et al. (2015) is the *screening statistic*. This screening statistic makes use of the screening set $\widehat{S}$ defined by:

$$\widehat{S} = \left\{ j \leq p : \frac{|\widehat{\theta}_j|}{\sqrt{\widehat{v}_j}} > \delta_{p,n} \right\},$$

where $\delta_{p,n}$ denotes a threshold sequence which is a function of $p$ and $n$. Hence, this screening set includes all components that have a pronounced signal. The screening statistic is then given by:

$$J_0 = \sqrt{p} \sum_{j \in \widehat{S}} \widehat{\theta}_j^2 \widehat{v}_j^{-1} = \sqrt{p} \sum_{j=1}^{p} \widehat{\theta}_j^2 \widehat{v}_j^{-1} \mathbb{I}_{j \in \widehat{S}},$$

where $\mathbb{I}_{j \in \widehat{S}}$ is an indicator variable which equals 1 if a component lies in $\widehat{S}$ and 0 otherwise. The screening statistic uses the screening set and screens out all components that do not have a pronounced signal. We can determine how pronounced this signal should be by changing the value of the threshold $\delta_{p,n}$. We want $\delta_{p,n}$ to be such that asymptotically under the null, no signal is pronounced enough and $J_0$ becomes zero. Here, Fan et al. (2015) suggests that $\delta_{p,n}$ should be chosen a bit higher than the noise level $\max_{j \leq p} \frac{|\widehat{\theta}_j - \theta_j|}{\sqrt{\widehat{v}_j}}$. Then it follows that under the null $H_0 : \boldsymbol{\theta} = 0$, $\widehat{S}$ becomes an empty set. Therefore, $J_0$ becomes equal to zero. However, note that choosing $\delta_{p,n}$ just higher than $\max_{j \leq p} \frac{|\widehat{\theta}_j - \theta_j|}{\sqrt{\widehat{v}_j}}$ is impossible in practice, since the true value of $\boldsymbol{\theta}$ is unknown. To make sure that the threshold is always larger than the noise level asymptotically, Fan et al. (2015) require that $\delta_{p,n} \to \infty$ as $n \to \infty$. A possible choice for this $\delta_{p,n}$ suggested in their paper is given by $\delta_{p,n} = \log(\log n)\sqrt{\log p}$. In Section 4.1, this choice will be discussed in more detail.

## 3.2 Verification of conditions

We will now check if the screening statistic indeed satisfies conditions (a), (b), and (c) to be a power enhancement component. Here, we follow the derivations from Fan et al. (2015).

The first condition is easily verified since we have a sum of non-negative terms so the screening statistic can never become negative.

Now condition (b) can be verified by giving a condition which the threshold $\delta_{p,n}$ should

satisfy. We want this threshold to be such that it dominates the maximum noise level. Written mathematically, we require under the null $H_0 : \boldsymbol{\theta} = \mathbf{0}$ that:

$$\mathbb{P}\left( \max_{j \leq p} \frac{|\widehat{\theta}_j - \theta_j|}{\sqrt{\widehat{v}_j}} < \delta_{p,n} \;\middle|\; H_0 \right) \to 1.$$

So, under the null, this maximum noise level will be smaller than $\delta_{p,n}$ with probability approaching 1 as $n \to \infty$. Using this, we find that:

$$\mathbb{P}(J_0 = 0 | H_0) = \mathbb{P}(\widehat{S} = \emptyset | H_0) = \mathbb{P}\left( \max_{j \leq p} \frac{|\widehat{\theta}_j - \theta_j|}{\sqrt{\widehat{v}_j}} \leq \delta_{p,n} \;\middle|\; H_0 \right) \to 1,$$

and hence condition (b) is also satisfied. So, the screening statistic satisfies the non-negativeness and the no-size-distortion property.

To show the last property, we define the following set suggested by Fan et al. (2015):

$$S(\boldsymbol{\theta}) = \left\{ j \leq p : \frac{|\theta_j|}{\sqrt{v_j}} > 3\delta_{p,n} \right\},$$

which depends on the vector $\boldsymbol{\theta}$. The idea of this set is the following: given some $\boldsymbol{\theta}$, this set contains the components which differ noticeably from zero. Note that we in general cannot derive which elements are in this set since we do not know $\boldsymbol{\theta}$ and $\boldsymbol{v}$. Furthermore, it follows by definition that $S(\mathbf{0}) = \emptyset$. Now, let's consider some $\boldsymbol{\theta} \neq \mathbf{0}$ for which $S(\boldsymbol{\theta})$ is not empty. It seems intuitive that the components which are in $S(\boldsymbol{\theta})$, will also be in $\widehat{S}$. Because of the multiplication of the threshold $\delta_{p,n}$ with 3 in $S(\boldsymbol{\theta})$, it seems likely that their estimate $\frac{|\widehat{\theta}_j|}{\widehat{v}_j}$ will be larger than $\delta_{p,n}$ and therefore, will be included in $\widehat{S}$. This intuition turns out to be true, as shown in Theorem 3.1 of the work of Fan et al. (2015), it holds for all $\boldsymbol{\theta} \in \mathbb{R}^p$ that $\mathbb{P}(S(\boldsymbol{\theta}) \subset \widehat{S} | \boldsymbol{\theta}) \to 1$ as $n \to \infty$. The technical details of this theorem can be found in the earlier manuscript Fan, Liao, and Yao (2014). This is a fairly strong asymptotic result: when just a few components differ noticeably from the null, we know that they will be included in the screening set $\widehat{S}$ with probability approaching 1.

Now let us use this result. Suppose that we are in some region of the alternative $H_1 : \boldsymbol{\theta} \neq 0$ where $S(\boldsymbol{\theta})$ is not empty. Then we find the following asymptotically:

$$\begin{aligned}
\mathbb{P}(J_0 > \sqrt{p} \mid S(\boldsymbol{\theta}) \neq \emptyset) &= \mathbb{P}\left( \sqrt{p} \sum_{j \in \widehat{S}} \frac{\widehat{\theta}_j^2}{\widehat{v}_j} > \sqrt{p} \;\middle|\; S(\boldsymbol{\theta}) \neq \emptyset \right) \\
&\overset{(1)}{\geq} \mathbb{P}\left( \sqrt{p} \sum_{j \in \widehat{S}} \delta_{p,n}^2 > \sqrt{p} \;\middle|\; S(\boldsymbol{\theta}) \neq \emptyset \right) \\
&= \mathbb{P}\left( \sum_{j \in \widehat{S}} \delta_{p,n}^2 > 1 \;\middle|\; S(\boldsymbol{\theta}) \neq \emptyset \right) \\
&\overset{(2)}{\to} 1,
\end{aligned}$$

where at (1) we use that $\widehat{S}$ is nonempty if $S(\boldsymbol{\theta})$ is nonempty with probability 1 asymptotically and that these squared components are at least as large as $\delta_{p,n}^2$. At (2) we use that $\delta_{p,n} \to \infty$

as $n \to \infty$. Hence, in this region where $S(\boldsymbol{\theta})$ is nonempty, $J_0$ will be larger than $\sqrt{p}$ with probability approaching 1, which shows the divergence of $J_0$. We can now conclude that the screening statistic indeed satisfies the properties to be a power enhancement component.

## 3.3 Null distribution of the test statistic

Now that we know how the power enhancement test statistic works, we also want to determine the distribution under the null. This is needed to determine the critical region. First, let us take a look at the initial test. As already stated before, we take a test here which already has correct asymptotic size. This means that we know the asymptotic distribution of its test statistic under the null. Written mathematically, we know that

$$J_1|H_0 \xrightarrow{\mathcal{D}} F, \quad \text{as} \quad n \to \infty,$$

where the $\mathcal{D}$ denotes convergence in distribution and $F$ is the limiting distribution. Now let $\alpha \in (0, 1)$ denote the significance level. Furthermore, we use $F_\alpha$ to denote the critical value. In this test, we have a right-tailed rejection region, so if the test statistic is larger than $F_\alpha$, then we reject the null. So, we know for this initial test that it has correct asymptotic size, which mathematically means that:

$$\lim_{n\to\infty} \mathbb{P}(J_1 > F_\alpha|H_0) = \alpha.$$

Now let us consider the null distribution of $J = J_0 + J_1$. As stated by property (b) of the power enhancement component, we know that $\mathbb{P}(J_0 = 0|H_0) \to 1$. Because of this, it is easily seen what is the distribution of $J$ under the null. First note that $\mathbb{P}(J_0 = 0|H_0) \to 1$ implies $J_0|H_0 \xrightarrow{\mathbb{P}} 0$. Then it follows by Slutsky's theorem that:

$$J_0|H_0 + J_1|H_0 \xrightarrow{\mathcal{D}} F.$$

Hence, the asymptotic null distribution of $J$ is also given by $F$. Now it follows that we can again use $F_\alpha$ as critical value for this test. Consequently, the asymptotic size of our test is known and given by:

$$\lim_{n\to\infty} \mathbb{P}(J > F_\alpha|H_0) = \alpha.$$

Since we only know the asymptotic size, we have no guarantees about the size of the test when we apply it in a finite sample. This will be further discussed in Section 4.1.

## 3.4 A sparse example

To finish this section about the technique itself, we will provide an example of an application. Note that our current choice of the power enhancement component, the screening statistic, especially improves the power against the sparse alternatives. To illustrate this, we will show what happens when we apply this test on the example of Section 2.4. In this example, we defined the $p(n)$-dimensional vector:

$$\boldsymbol{\theta}_n = (\log p(n), 0, \ldots, 0).$$

We will again consider the case in which this vector is the mean of a multivariate normal distribution. As was shown, the conventional Wald test has low power against this alternative

and this power eventually even converges to the significance level. To show what happens when we use the power enhancement test against this alternative, we will use the set $S(\boldsymbol{\theta}) = \left\{ j \leq p : \frac{|\theta_j|}{\sqrt{v_j}} > 3\delta_{p,n} \right\}$. Since all components after the first component of $\boldsymbol{\theta}_n$ are zero, they will not be included in this set. However, for the first component we have the following condition:

$$\frac{|\theta_1|}{\sqrt{v_1}} > 3\delta_{p,n} \iff \log p(n) > 3\log(\log n)\sqrt{\log p(n)} \iff n\log 2 > \log(\log n)\sqrt{n\log 2},$$

which holds for all $n \in \mathbb{N}$. Therefore, we know for this alternative $\boldsymbol{\theta}_n$ that $S(\boldsymbol{\theta}_n) = \{1\}$. As shown in Section 3.2, we now know that $J_0$ will diverge with $p$ and therefore, boost the power against this alternative $\boldsymbol{\theta}_n$.

The idea of the set $S(\boldsymbol{\theta})$ is that it will capture the sparse signals. However, it is important to notice here that this does not always have to be in line with the sparsity measures discussed in Section 2.1. That is, some vectors may be equally sparse according to a sparsity measure, but for only one of them $S(\boldsymbol{\theta})$ is empty. To illustrate this, we consider the following example:

$$\widetilde{\boldsymbol{\theta}}_n = (\sqrt{\log p(n)}, 0, \ldots, 0).$$

According to all sparsity measures discussed in Section 2.1, this vector is just as sparse as $\boldsymbol{\theta}_n$. However, when we plug this vector into $S(\boldsymbol{\theta})$, we get the condition:

$$\frac{|\theta_1|}{\sqrt{v_1}} > 3\delta_{p,n} \iff \sqrt{n\log 2} > \log(\log n)\sqrt{n\log 2},$$

which only holds for $n \leq 15$. Therefore, as we start increasing $n$, we find that $S(\widetilde{\boldsymbol{\theta}}_n)$ will be empty. This reveals that the practitioner should be cautious and that it is possible that the power enhancement technique performs differently for vectors of the same sparsity level.

In our analysis, we want to compare the performance of the technique for different sparsity levels and therefore we also need to choose which sparsity measure we are going to use. To easily change the sparsity of the alternatives, we will restrict ourselves to vectors only consisting of two distinct values. These values will be zero and a constant which is yet to be determined. Because the $\ell_0-$norm is a good measure sparsity in this setting, we will use this one. The problem with the Hoyer index is that it is not suitable to compare vectors of different lengths. Moreover, the Gini index could also be used here but since the $\ell_0$-norm performs just as well in this case, we will use the simpler measure. Also, note that the $\ell_0$-norm is a more natural interpretation of the $S(\boldsymbol{\theta})$ set, there are few components that have a large signal and only those should be counted.

# 4 Problem formulation

The power enhancement technique seems a very promising technique. Nevertheless, there might be some pitfalls when applying this technique. In the literature, these pitfalls have not been researched thoroughly yet. This section discusses what these pitfalls might be.

## 4.1 The problem in practice

As shown in the work of Kock and Preinerstorfer (2019), the power enhancement technique has strong asymptotic results. In their paper, they showed that for all sufficiently slowly increasing growth rates of the dimension of the parameter vector, every test with asymptotic size less than one, will obtain more power. These tests do not lose any size asymptotically, so it seems like this power is almost 'for free'. However, when we have a finite sample this cannot be the case. In this case, there is a positive probability under the null that $J_0 \neq 0$. Consequently, in a finite sample, the critical value of $J_0 + J_1$ needs to be strictly larger than than the critical value of $J_1$. So, by using the critical value of $J_1$ for $J_0 + J_1$, it is guaranteed that we get size distortion. Now, the crucial question which this study aims to answer is, how large is this size distortion? If this size distortion is very large, the power enhancement technique might not be that promising in practice as it seems to be asymptotically. However, it could also be that this size distortion is fairly small. Then the practitioner might willingly give up a little bit of size to obtain a large increase in power.

The size distortion goes hand in hand with the threshold sequence $\delta_{p,n}$. Recall that the screening set $\widehat{S}$ is given by:

$$\widehat{S} = \left\{ j \leq p : \frac{|\widehat{\theta}_j|}{\sqrt{\widehat{v}_j}} > \delta_{p,n} \right\}.$$

So, a component is included in this set if $\frac{|\widehat{\theta}_j|}{\sqrt{\widehat{v}_j}}$ is larger than the threshold $\delta_{p,n}$. The choice of an appropriate threshold sequence is crucial for this technique to work. A threshold that is too strong will barely increase the power. However, a threshold that is too soft will give a large size distortion.

What we want for the threshold sequence is that it is slightly larger than $\max_{j \leq p} \frac{|\widehat{\theta}_j - \theta_j|}{\sqrt{\widehat{v}_j}}$. In this way, it will not pick up the noise but it will pick up the significant signals. Note that this maximum will always be non-decreasing when we keep increasing the value of $p$. So to choose an appropriate sequence for $\delta_{p,n}$, we want to know how $\max_{j \leq p} \frac{|\widehat{\theta}_j - \theta_j|}{\sqrt{\widehat{v}_j}}$ approximately increases when $p$ increases. It follows from large deviation theory, under some conditions, that typically $\max_{j \leq p} \frac{|\widehat{\theta}_j - \theta_j|}{\sqrt{\widehat{v}_j}} = O_p(\sqrt{\log p})$. The proof of this is out of the scope of this study.

It also seems intuitive that the threshold should be growing in $n$. Consider the example in which we are testing a hypothesis on a vector which is the mean of a multivariate normal distribution with the identity matrix as its covariance matrix. The condition to be included in $\widehat{S}$ then becomes $\sqrt{n}|\widehat{\theta}_j| > \delta_{p,n}$. Notice that, under the null, the distribution of the left-hand side of this equation does not depend on $n$. Now, if we would fix $p$ and keep increasing $n$, we would still like to see that no components are included asymptotically. To make sure this happens, we also want $\delta_{p,n}$ to grow in $n$.

Hence, we want that $\delta_{p,n}$ grows in either $p$ and $n$. In their paper, Fan et al. (2015) suggest the following threshold sequence:

$$\delta_{p,n} = \log(\log n)\sqrt{\log p}.$$

This sequence is indeed growing in either $p$ and $n$. Nevertheless, it remains unclear how the $\log(\log n)$ term is chosen. This term is also used in the work of Fan and Tang (2012)

where it is used to optimize the generalized information criterion.

Since Fan et al. (2015) only give asymptotic arguments for this choice of threshold, it is interesting to see how sensitive the test is to this threshold in a finite sample. Furthermore, when looking into the supplementary material provided by Fan et al. (2015), it can be found that there are some variations in the use of $\delta_{p,n} = \log(\log n)\sqrt{\log p}$ in the simulations. Instead of just using $\delta_{p,n}$, the simulations were done by using $\sqrt{1.5}\delta_{p,n}$, $1.06\delta_{p,n}$, $\delta_{p,n}$, and $0.9\delta_{p,n}$ as thresholds. Of course, this constant in front does not matter asymptotically. However, it might have serious consequences in a finite sample.

To illustrate this, consider the following example. Suppose that $n = 100$ and that the constant is $\sqrt{1.5}$. Now, for the threshold, we have $\sqrt{1.5}\delta_{p,100} = \log((\log 100)^{\sqrt{1.5}})\sqrt{\log p} \approx \log 6.49\sqrt{\log p}$. This yields the same threshold as just using 1 as a constant, taking $n = e^{6.49} \approx 659$, and keeping $p$ the same. This clearly demonstrates that a small change in this constant can have a large effect.

A side note about this threshold $\delta_{p,n}$ is that it becomes negative for $n = 1$ and $n = 2$. In that case, every component will be in $\widehat{S}$ and it does not make sense to use the power enhancement technique. From now on we will assume that $n > 2$ without mention.

# 5 Theoretical analysis

To obtain some more insights into the practical use of the technique, we first perform a theoretical analysis. In this analysis, we derive the size and power of the technique under some simplifying assumptions which we make along the way. This expression for the size and power becomes computationally too difficult when the number of parameters $p$ becomes large. Therefore, we also approximate this expression numerically. Moreover, we derive an upper and lower bound which are computationally even more pleasant. From these bounds, we also obtain some intuitive insights. Finally, we show how the size and power depend on the number of parameters $p$ and the number of observations $n$. Proofs of lemmas and propositions given in this section will appear in the appendix.

## 5.1 General setup

We define the following data generating process:

$$\boldsymbol{X} = \boldsymbol{\iota}_n \boldsymbol{\theta}^T + \boldsymbol{E},$$

where $\boldsymbol{\iota}_n$ is an $n$-dimensional vector of ones and $\boldsymbol{E}$ is a random matrix of dimension $n \times p$. Hence, $\boldsymbol{X}$ is a matrix of $n$ rows and $p$ columns where each column $j$ contains $\theta_j$ plus some added noise. We do assume independence of all elements and that the noise has mean zero.

To apply the power enhancement technique, we first need to get an estimate of $\boldsymbol{\theta}$. In this case, that means taking the sample means of the columns of $\boldsymbol{X}$. This yields a consistent estimator of $\boldsymbol{\theta}$ since the noise has mean zero. Written in matrix notation, we have:

$$\widehat{\boldsymbol{\theta}} = \frac{1}{n}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{\iota}_n.$$

Now, consider the case in which we use the Wald test to test the hypothesis $H_0 : \boldsymbol{\theta} = \boldsymbol{0}$. We denote the test statistic of the Wald test by $J_1$. Moreover, the significance level is denoted by $\alpha$ and the critical value by $F_\alpha$. Hence, we reject $H_0 : \boldsymbol{\theta} = \boldsymbol{0}$ if $J_1 > F_\alpha$. We

are now interested in what happens when we apply the power enhancement technique on this test. We will use the screening statistic discussed in Section 3.1 as power enhancement component and we will again denote it by $J_0$. For both the size and the power, we are interested in the probability of rejection which is given by:

$$\mathbb{P}(J_0 + J_1 > F_\alpha). \tag{1}$$

Recall that the screening set is given by:

$$\widehat{S} = \{j \leq p : \frac{|\widehat{\theta}_j|}{\sqrt{\widehat{v}_j}} > \delta_{p,n}\},$$

where $\widehat{v}_j$ denotes the estimated asymptotic variance. We will now rewrite equation (1). To do this, we condition on the screening set $\widehat{S}$ since it contains some important information to determine $J_0$.

Let us use $A$ to denote the event $J_0 + J_1 > F_\alpha$ and let $\mathbb{I}_A$ denote the indicator variable which equals 1 if $A$ happens to be true. Now, using the Fundamental Bridge and the Tower Rule, we find that:

$$\mathbb{P}(A) = \mathbb{E}[\mathbb{I}_A] = \mathbb{E}[\mathbb{E}[\mathbb{I}_A \mid \widehat{S}]] = \mathbb{E}[\mathbb{P}(A \mid \widehat{S})] = \mathbb{E}[\mathbb{P}(J_0 + J_1 > F_\alpha \mid \widehat{S})]. \tag{2}$$

To denote the different forms that this screening set may take more rigorously, we introduce the following notation. Let $S = \{1, \ldots, p\}$ be the set of all components. The power set of this set is then defined as all possible subsets of this set, including the empty set and the set itself. Hence, this power set shows all possible forms that $\widehat{S}$ can take. We will denote this power set by $\mathcal{P}(S)$, which contains all subsets of $S$. We can now use this to write the expectation in equation (2) out:

$$\mathbb{E}[\mathbb{P}(J_0 + J_1 > F_\alpha \mid \widehat{S})] = \sum_{s \in \mathcal{P}(S)} \mathbb{P}(J_0 + J_1 > F_\alpha \mid \widehat{S} = s)\mathbb{P}(\widehat{S} = s). \tag{3}$$

We will now explain how this expression converges to the significance level under the null. Under the null, as $p$ and $n$ start increasing, the probability $\mathbb{P}(\widehat{S} = \emptyset)$ becomes larger and larger and converges to 1. Therefore, in the limit, only the first term of the summation is non-zero. This term is given by $\mathbb{P}(J_0 + J_1 > F_\alpha \mid \widehat{S} = \emptyset)\mathbb{P}(\widehat{S} = \emptyset)$. For the conditional probability here, we know that:

$$\mathbb{P}(J_0 + J_1 > F_\alpha \mid \widehat{S} = \emptyset) = \frac{\mathbb{P}(J_0 + J_1 > F_\alpha \cap \widehat{S} = \emptyset)}{\mathbb{P}(\widehat{S} = \emptyset)} = \frac{\mathbb{P}(J_1 > F_\alpha \cap \widehat{S} = \emptyset)}{\mathbb{P}(\widehat{S} = \emptyset)}.$$

In the numerator, we have the probability of the events $J_1 > F_\alpha$ and $\widehat{S} = \emptyset$ happening simultaneously. However, the event $\widehat{S} = \emptyset$ tells us that all components are relatively small. Therefore, this event has a negative effect on the probability that $J_1 > F_\alpha$. This tells us that $\mathbb{P}(J_1 > F_\alpha \cap \widehat{S} = \emptyset) < \mathbb{P}(J_1 > F_\alpha)\mathbb{P}(\widehat{S} = \emptyset)$. Because of this, we find that:

$$\mathbb{P}(J_0 + J_1 > F_\alpha \mid \widehat{S} = \emptyset) = \frac{\mathbb{P}(J_1 > F_\alpha \cap \widehat{S} = \emptyset)}{\mathbb{P}(\widehat{S} = \emptyset)} < \frac{\mathbb{P}(J_1 > F_\alpha)\mathbb{P}(\widehat{S} = \emptyset)}{\mathbb{P}(\widehat{S} = \emptyset)} = \mathbb{P}(J_1 > F_\alpha).$$

Hence, the above conditional probability is smaller than the significance level $\alpha = \mathbb{P}(J_1 > F_\alpha)$. However, as $p$ and $n$ start increasing, the threshold also increases and therefore the

probability of intersection $\mathbb{P}(J_1 > F_\alpha \cap \widehat{S} = \emptyset)$ also rises. Eventually, the threshold will become so large that this term will converge to the significance level $\alpha$. From this, we conclude that the expression (3) will converge to the significance level under the null.

It is difficult to gain further insights into this expression. Therefore, we will look at a special case.

## 5.2 Assumption of identicality

By this assumption, we mean that all errors in random matrix $\boldsymbol{E}$ are identically distributed. Recall that we already assumed that the errors are independently distributed so we now have i.i.d. errors. Furthermore, we also assume that $\theta_1 = \ldots = \theta_p$. We make this assumption such that all components of vector $\widehat{\theta}$ are also independently and identically distributed. This assumption means that $\boldsymbol{\theta}$ can be the zero vector or a scalar multiple of $\boldsymbol{\iota}_p$. Hence, $\boldsymbol{\theta}$ takes on the value under the null or it takes on the value of a very dense alternative.

This assumption gives a simple approach to compute the distribution of $\widehat{S}$. In particular, the probability that component $j$ is included in $\widehat{S}$, $\mathbb{P}\left(\frac{|\widehat{\theta}_j|}{\sqrt{\widehat{v}_j}} > \delta_{p,n}\right)$, is equal for all $j$. So, the number of elements in $\widehat{S}$, denoted by $|\widehat{S}|$, is binomially distributed with parameters $p$ and $q$, where $q$ is the inclusion probability of a single component defined above. Instead of summing over the powersets as in (3), this allows us to sum over the number of elements included:

$$\sum_{s \in \mathcal{P}(S)} \mathbb{P}(J_0 + J_1 > F_\alpha \mid \widehat{S} = s)\mathbb{P}(\widehat{S} = s) = \sum_{s=0}^{p} \mathbb{P}(J_0 + J_1 > F_\alpha \mid |\widehat{S}| = s)\mathbb{P}(|\widehat{S}| = s). \quad (4)$$

Now, for the conditional probability $\mathbb{P}(J_0 + J_1 > F_\alpha \mid |\widehat{S}| = s)$, we do not have to think about which elements are included in $\widehat{S}$ but we can just focus on the number of elements in it. Nonetheless, it is still difficult to analyze this expression. For this reason, we will make another assumption.

## 5.3 Assumption of normality

We now also assume that all errors in the random matrix $\boldsymbol{E}$ are standard normally distributed. Since any linear combination of normally distributed random variables is also normally distributed, we get that:

$$\widehat{\boldsymbol{\theta}} = \frac{1}{n}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{\iota}_n \sim \mathcal{N}\left(\boldsymbol{\theta}, \frac{1}{n}\boldsymbol{I}\right).$$

Using this assumption we can simplify (4). Firstly, we now know how to compute $q$. Standard normality tells us that $\widehat{v}_j = \frac{1}{n}$ and therefore we get that $q = \mathbb{P}(\sqrt{n}|\widehat{\theta}_1| > \delta_{p,n})$. Now, also note that normality tells us that $\sqrt{n}\widehat{\theta}_1 \sim \mathcal{N}(\sqrt{n}\theta_1, 1)$. From this, it follows that $n\widehat{\theta}_1^2 \sim \chi^2_{1,n\theta_1^2}$, where the first subscript denotes the degrees of freedom and the second subscript denotes the non-centrality parameter. Note we are now able to compute parameter $q$ using that:

$$q = \mathbb{P}(\sqrt{n}|\widehat{\theta}_1| > \delta_{p,n}) = \mathbb{P}(n\widehat{\theta}_1^2 > \delta_{p,n}^2),$$

and hence, the probability $\mathbb{P}(|\widehat{S}| = s)$ in (4) is known for all $s$.

We are now left with computing the conditional probability in (4) which is given by:

$$\mathbb{P}(J_0 + J_1 > F_\alpha \mid |\widehat{S}| = s). \tag{5}$$

To compute this probability, we need to know the distribution of $J_0 = \sqrt{p}n \sum_{j \in \widehat{S}} \widehat{\theta}_j^2$. Here, we can again use our assumption of normality which tells us that that for any component $j$, we have that $n\widehat{\theta}_j^2 \sim \chi^2_{1,n\theta_j^2}$. This is useful since $J_0$ is the sum of these $n\widehat{\theta}_j^2$ terms multiplied by $\sqrt{p}$. Furthermore, we also know something about the magnitude of these terms. That is, under standard normality, the condition for a component to be included in $\widehat{S}$ is $\sqrt{n}|\widehat{\theta}_j| > \delta_{p,n}$ which is equivalent to $n\widehat{\theta}_j^2 > \delta_{p,n}^2$. Hence, the condition that $|\widehat{S}| = s$ not only tells us how many components are included in $J_0$ but it also gives us a condition about the magnitude of these components. Without loss of generality, let us assume that $\widehat{S}$ contains the first $s$ components. Generality is not lost as $J_0$ only depends on the number of elements in $\widehat{S}$ due to the i.i.d. assumption. This is summarized by the following lemma.

**Lemma 1.** *For any number $s \in \{0, \dots, p\}$ of components in $\widehat{S}$, the conditional probability of rejection is equal to the conditional probability with the first $s$ components included in $\widehat{S}$, that is:*

$$\mathbb{P}(J > F_\alpha \mid |\widehat{S}| = s) = \mathbb{P}(J > F_\alpha \mid n\widehat{\theta}_1^2 > \delta_{p,n}^2, \dots, n\widehat{\theta}_s^2 > \delta_{p,n}^2, n\widehat{\theta}_{s+1}^2 < \delta_{p,n}^2, \dots, n\widehat{\theta}_p^2 < \delta_{p,n}^2).$$

Now, using this lemma, equation (5) can be rewritten as:

$$\begin{aligned}\mathbb{P}(J_0 + J_1 > F_\alpha \mid n\widehat{\theta}_1^2 > \delta_{p,n}^2, \dots, n\widehat{\theta}_s^2 > \delta_{p,n}^2, \\ n\widehat{\theta}_{s+1}^2 < \delta_{p,n}^2, \dots, n\widehat{\theta}_p^2 < \delta_{p,n}^2) = \mathbb{P}(J_0 + J_1 > F_\alpha \mid C_s),\end{aligned} \tag{6}$$

where $C_s$ will be used to replace the conditions for the rest of the analysis for the sake of notational simplicity.

Recall that $J_1$ is the test statistic of the Wald test. The Wald test statistic is quite useful here to determine the probability since it also contains the $n\widehat{\theta}_j^2$ terms. In this setting, the Wald statistic is given by:

$$J_1 = \widehat{\boldsymbol{\theta}}^{\boldsymbol{T}} \mathrm{Var}(\widehat{\boldsymbol{\theta}})^{-1} \widehat{\boldsymbol{\theta}} = n\widehat{\boldsymbol{\theta}}^{\boldsymbol{T}} \widehat{\boldsymbol{\theta}} = n \sum_{j=1}^{p} \widehat{\theta}_j^2,$$

where we used our assumption that $\mathrm{Var}(\widehat{\boldsymbol{\theta}}) = \frac{1}{n}\boldsymbol{I}$. Now plugging $J_0$ and $J_1$ into (6), gives us that:

$$\mathbb{P}(J_0 + J_1 > F_\alpha \mid C_s) = \mathbb{P}\left(\sqrt{p}n \sum_{j=1}^{s} \widehat{\theta}_j^2 + n \sum_{j=1}^{p} \widehat{\theta}_j^2 > F_\alpha \mid C_s\right). \tag{7}$$

In this expression, we do know the individual distributions of the $n\widehat{\theta}_j^2$ terms and since they are independent, the joint distribution is straightforward. However, because of the condition

$C_s$, we need the conditional joint density of the random variables $(n\widehat{\theta}_1^2, \ldots, n\widehat{\theta}_p^2 \mid C_s)$ to compute expression (7). This is given by:

$$
\begin{aligned}
f_{n\widehat{\theta}_1^2,\ldots,n\widehat{\theta}_p^2 \mid C_s}(x_1,\ldots,x_p) &= \frac{f_{n\widehat{\theta}_1^2,\ldots,n\widehat{\theta}_p^2}(x_1,\ldots,x_p)}{\mathbb{P}(n\widehat{\theta}_1^2 > \delta_{p,n}^2, \ldots, n\widehat{\theta}_s^2 > \delta_{p,n}^2, n\widehat{\theta}_{s+1}^2 < \delta_{p,n}^2, \ldots, n\widehat{\theta}_p^2 < \delta_{p,n}^2)} \\
&\overset{\text{i.i.d.}}{=} \frac{\prod_{j=1}^p f_{n\widehat{\theta}_j^2}(x_j)}{\mathbb{P}\left(n\widehat{\theta}_1^2 > \delta_{p,n}^2\right)^s \mathbb{P}\left(n\widehat{\theta}_1^2 < \delta_{p,n}^2\right)^{p-s}} \\
&= \frac{\prod_{j=1}^p f_{n\widehat{\theta}_j^2}(x_j)}{\overline{F}_{n\widehat{\theta}_1^2}(\delta_{p,n}^2)^s F_{n\widehat{\theta}_1^2}(\delta_{p,n}^2)^{p-s}}, \quad \text{for } (x_1,\ldots,x_p) \in (\delta_{p,n}^2, \infty)^s \times (0, \delta_{p,n}^2)^{p-s} \\
&\qquad\qquad \text{and } 0 \text{ otherwise.}
\end{aligned}
$$

Here, $\overline{F}_{n\widehat{\theta}_1^2}$ is used to denote the survivor function of $n\widehat{\theta}_1^2$. Now that we have this joint density we can compute the conditional probability from equation (7). For notational simplicity, we introduce the indicator function $g(\widehat{\boldsymbol{\theta}}) = \mathbb{I}_{\sqrt{p}n\sum_{j=1}^s \widehat{\theta}_j^2 + n\sum_{j=1}^p \widehat{\theta}_j^2 > F_\alpha}$. This gives us our final expression for the conditional probability given in (5):

$$
\begin{aligned}
\mathbb{P}(J_0 + J_1 > F_\alpha \mid |\widehat{S}| = s) &= \mathbb{P}\left(\sqrt{p}n\sum_{j=1}^s \widehat{\theta}_j^2 + n\sum_{j=1}^p \widehat{\theta}_j^2 > F_\alpha \mid C_s\right) \\
&= \mathbb{E}\left[\mathbb{I}_{\sqrt{p}n\sum_{j=1}^s \widehat{\theta}_j^2 + n\sum_{j=1}^p \widehat{\theta}_j^2 > F_\alpha} \mid C_s\right] \\
&= \mathbb{E}\left[g(\widehat{\boldsymbol{\theta}}) \mid C_s\right] \\
&= \int_{\delta_{p,n}^2}^\infty \cdots \int_{\delta_{p,n}^2}^\infty \int_0^{\delta_{p,n}^2} \cdots \int_0^{\delta_{p,n}^2} g(\widehat{\boldsymbol{\theta}})\ f_{n\widehat{\theta}_1^2,\ldots,n\widehat{\theta}_p^2 \mid C}(x_1,\ldots,x_p)\ dx_1 \ldots dx_p \\
&= \int_{\delta_{p,n}^2}^\infty \cdots \int_{\delta_{p,n}^2}^\infty \int_0^{\delta_{p,n}^2} \cdots \int_0^{\delta_{p,n}^2} g(\widehat{\boldsymbol{\theta}})\ \frac{\prod_{j=1}^p f_{n\widehat{\theta}_j^2}(x_j)}{\overline{F}_{n\widehat{\theta}_1^2}(\delta_{p,n}^2)^s F_{n\widehat{\theta}_1^2}(\delta_{p,n}^2)^{p-s}}\ dx_1 \ldots dx_p, \qquad (8)
\end{aligned}
$$

where we have $s$ integrals on the interval $(\delta_{p,n}^2, \infty)$ and $p - s$ integrals on the interval $(0, \delta_{p,n}^2)$. We were unable to simplify this expression. Therefore, we resort to numerical integration to compute this quantity. This expression does also immediately show the curse of dimensionality. In case we want to take a look at $p = 100$ for example, we would need to compute a 100-dimensional integral which is computationally very demanding. Nonetheless, we can compute the exact value in case we take a small $p$. In the next subsection, we will demonstrate this under the null hypothesis.

## 5.4 Exact size of the test

By using the expression derived earlier, we can now compute the size of the test by plugging in $\boldsymbol{\theta} = \mathbf{0}$. In Figure 5.1, this is done for $p = 2$. We computed this size for $n = 1, \ldots, 100$.

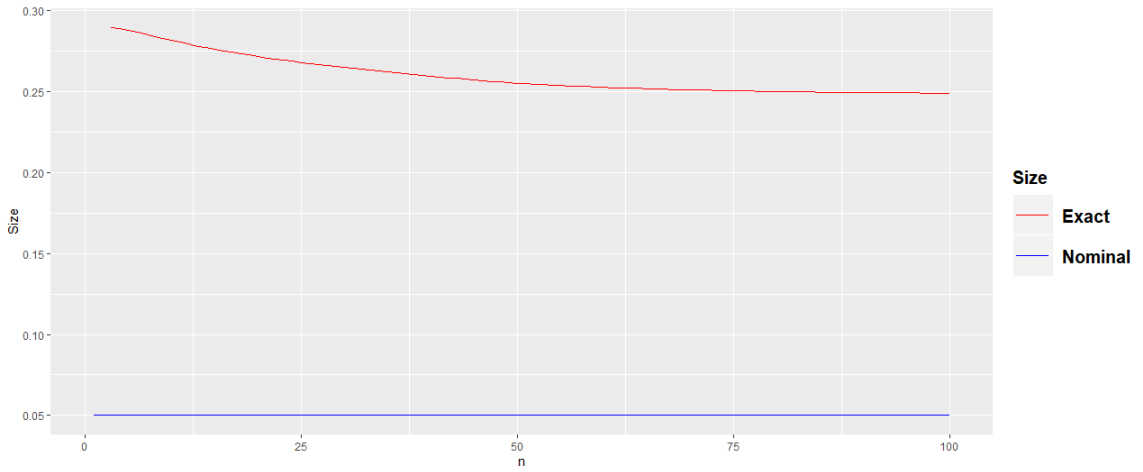


Figure 5.1: Actual and nominal size of the test in case $p = 2$

This figure exposes some serious size distortion. Furthermore, we observe that the size distortion decreases with $n$, which is also as expected. However, this happens at a very slow rate. This is due to the slow rate of divergence of $\delta_{p,n}$ in $n$. Unfortunately, for larger values of $p$ numerical integration becomes computationally prohibitive. For this reason, we resort to a method that efficiently approximates these high-dimensional integrals.

## 5.5 Monte Carlo integration

In this subsection, we exhibit the use of Monte Carlo integration to approximate the high-dimensional integral in (8). The idea of this method is the following. We need to integrate a function $\boldsymbol{f}(n\widehat{\theta}_1^2, \ldots, n\widehat{\theta}_p^2)$, which is the integrand in equation (8), over the space $(\delta_{p,n}^2, \infty)^s \times (0, \delta_{p,n}^2)^{p-s}$. This space is a subset of the space $(0, \infty)^p$ and on this larger space we know the joint density of $(n\widehat{\theta}_1^2, \ldots, n\widehat{\theta}_p^2)$, which is the product of $\chi^2$ distributions. This means that we know the distribution on some space and we want to integrate over a subset of that space. Let us use $\boldsymbol{p}(\cdot)$ to denote the product $\chi^2$ distribution and let $\bar{\boldsymbol{x}}_1, \ldots, \bar{\boldsymbol{x}}_N$ be $N$ samples from this distribution. Then, the importance sampling algorithm tells us that we can perform Monte Carlo integration using the expression:

$$Q_N = \frac{1}{N} \sum_{i=1}^{N} \frac{\boldsymbol{f}(\bar{\boldsymbol{x}}_i)}{\boldsymbol{p}(\bar{\boldsymbol{x}}_i)}.$$

Intuitively, this tells us that when a sample $\bar{\boldsymbol{x}}_i$ has a relatively high likelihood of occurring (so $\boldsymbol{p}(\bar{\boldsymbol{x}}_i)$ is high) compared to another sample $\bar{\boldsymbol{x}}_j$, we put less weight on this sample. This way we get an equal weighting over all points in the space and then the integral can be calculated as the average of these points evaluated at $\boldsymbol{f}$. Moreover, the law of large numbers ensures that

$$\lim_{N \to \infty} Q_N = I,$$

where $I$ is the integral in equation (8). To examine the quality of this approximation, we will first compare it with the exact case where $p = 2$. This can be seen in Figure 5.2. For the Monte Carlo approximation we used $N = 30,000$ here.

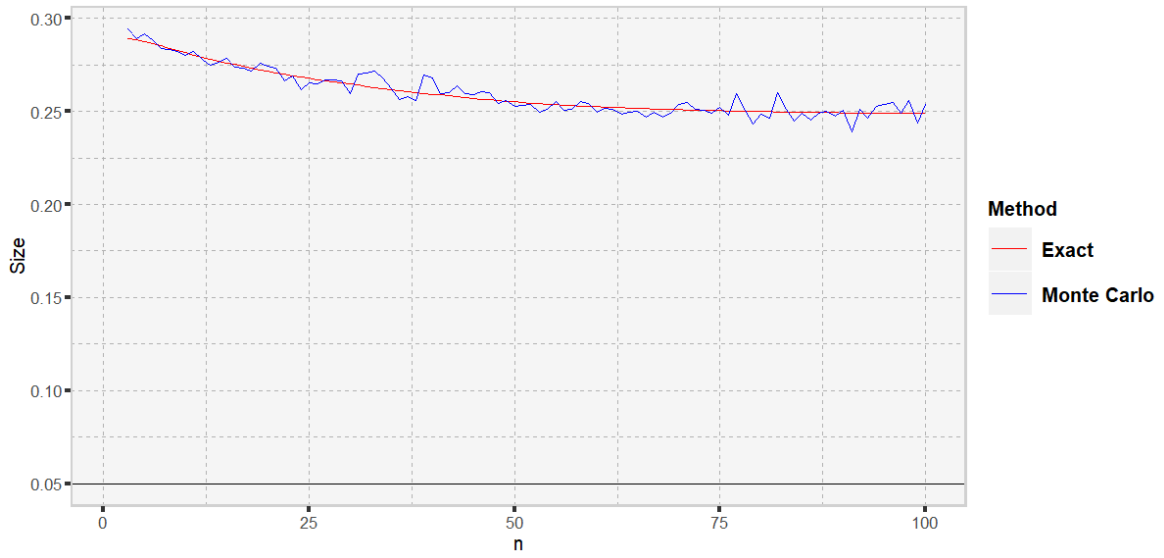


Figure 5.2: Exact size and its Monte Carlo approximation for $p = 2$

This approximation seems to be quite accurate and therefore, we will also use it to approximate the size distortion for a higher value of $p$. These approximations are shown in Figure 5.3.

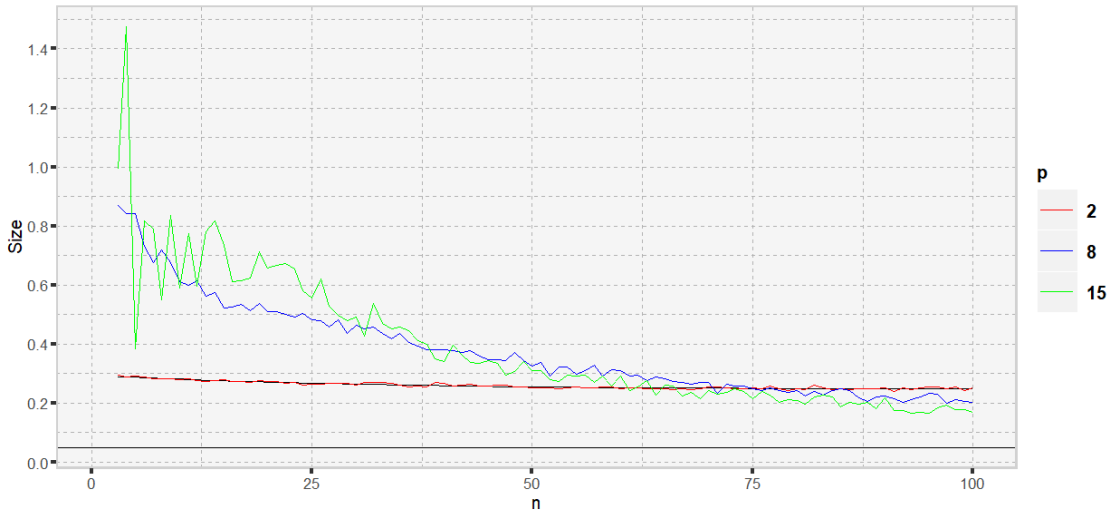


Figure 5.3: Monte Carlo approximation for multiple dimensions

Note that the approximation for $p = 15$ is not entirely accurate since it also gives an actual size which is larger than 1. This is due to the high variance of the approximations for small values of $n$. Nevertheless, this plot gives us some interesting insights. We again see that the size is decreasing with $n$, this happens for all values of $p$. When we start changing $p$, there does not seem to be such a monotonic effect. For small values of $n$, the size is clearly increasing with $p$. However, when we take a large value of $n$, the size is actually decreasing with $p$. This is an interesting observation and we will look further into this in Section 5.8.

Although this method of Monte Carlo integration makes it easier to approximate high-dimensional integrals, it still becomes computationally very demanding when we start increasing $p$. As can be seen in Figure 5.3, the approximation for $p = 15$ is already pretty

rough. It also took already about 15 minutes on a laptop with an Intel Core i7 processor to compute this approximation. Since we are interested in the performance of the power enhancement technique for larger values of $p$, like $p = 100$, we will construct an upper and lower bound which are a lot easier to compute. Before we do this, we will first decompose the sum from equation (4) and examine the effect of each summand separately. In this way, we obtain more insights into the behaviour of this sum and this can help us constructing an upper and lower bound.

## 5.6 Decomposition of effects

The i.i.d. assumption allowed us to write the probability of rejection as a sum of terms:

$$\sum_{s=0}^{p} \mathbb{P}(J_0 + J_1 > F_\alpha \mid |\widehat{S}| = s)\mathbb{P}(|\widehat{S}| = s).$$

It might be the case that the values of some of these summands are negligible. If this would be the case, we could skip these terms when computing the lower bound.

In the sum given above, the computational difficulty comes from the conditional probabilities. Notice that these conditional probabilities are weighted by the binomial probabilities from $|\widehat{S}|$. This binomial distribution has probability parameter $q = \mathbb{P}(\sqrt{n}|\widehat{\theta}_1| > \delta_{p,n})$. Under the null, $q$ becomes smaller as $p$ and $n$ increase. Therefore, the later summands of the sum are weighted with decreasingly small weights. We will now consider how these effects interact with each other under the null. In Figure 5.4, this has been done for $p = 2$.

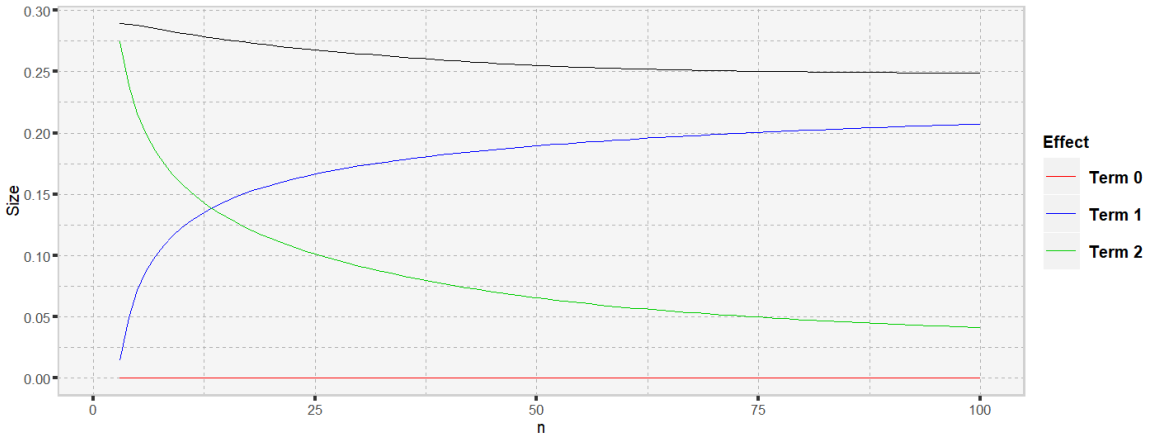


Figure 5.4: Decomposition of effects for $p = 2$ under the null.

When $n$ is still small, the total effect is largely due to the last term. This is explained by the effect that the binomial probability $q$ is very large and therefore, it is very likely that all components will be in $\widehat{S}$. When $n$ is large, this is not the case anymore. Here you might expect that term 0 would be dominating. However, this term is still equal to 0. This can be explained by taking a better look at the conditional probability:

$$\begin{aligned}\mathbb{P}(J_0 + J_1 > F_\alpha \mid |\widehat{S}| = 0) &= \mathbb{P}(J_1 > F_\alpha \mid |\widehat{S}| = 0) \\ &= \mathbb{P}(n\widehat{\theta}_1^2 + n\widehat{\theta}_2^2 > F_\alpha \mid n\widehat{\theta}_1^2 < \delta_{p,n}^2, n\widehat{\theta}_2^2 < \delta_{p,n}^2). \qquad (9)\end{aligned}$$

For small values of $p$ and $n$, the value of $\delta_{p,n}$ is so small that the condition that $n\widehat{\theta}_1^2 < \delta_{p,n}^2$ and $n\widehat{\theta}_2^2 < \delta_{p,n}^2$, is strong enough such that the event $n\widehat{\theta}_1^2 + n\widehat{\theta}_2^2 > F_\alpha$ cannot happen. Notice that as we keep increasing $n$, this will change at some point and this term will converge to the significance level.

It is also interesting to see how this decomposition of effects carries over to larger values of $p$ and $n$. Therefore, we have used the method of Monte Carlo to approximate these effects for $p = 10$. This decomposition can be found in Figure 5.5. Also, notice that $n$ goes to 1000 in this figure.

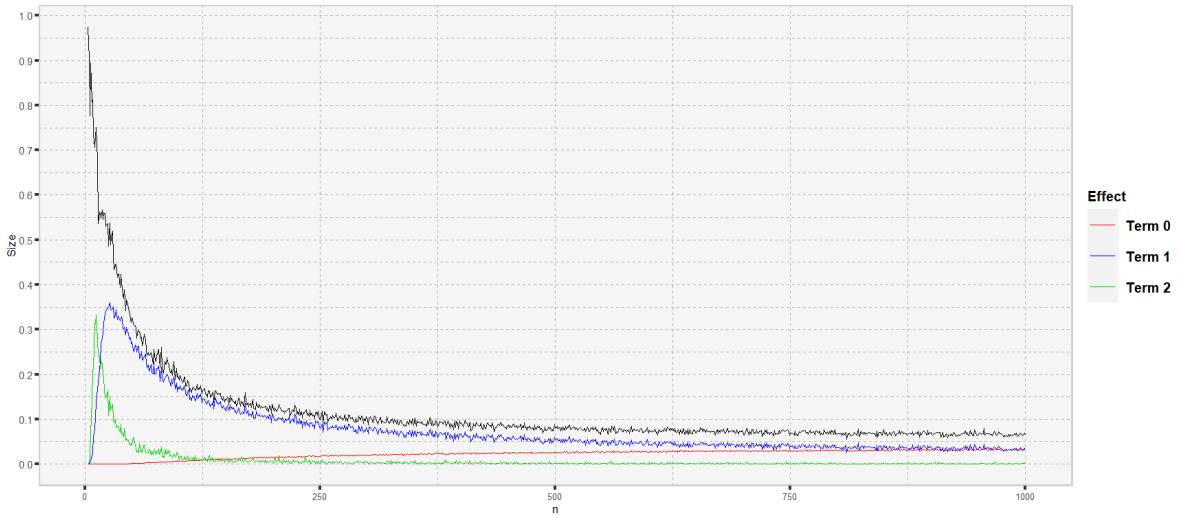


Figure 5.5: Decomposition of effects for $p = 10$ under the null.

In this figure, we have only shown the first three effects, while there are of course more since $p = 10$. However, it can be seen that these first three effects dominate rather quickly. We also observe that term 0 converges from below to the significance level as also explained at the end of Section 5.1. This convergence still happens at a very slow rate which can be explained by the slow divergence of $\delta_{p,n}$ in $n$. This demonstrates that the test is very dependent on the choice of the threshold $\delta_{p,n}$.

## 5.7 Upper and lower bound for the size

In this subsection, we will derive an efficiently computable upper and lower bound for the size. As we saw in the previous section that the first terms in (4) quickly start to dominate, we consider eliminating the later summands in these bounds. It is essential to notice here that it is only guaranteed under the null that the later summands become negligible. Under the alternative, it also might be the case that actually the first terms become negligible. Because of this, we restrict our attention to the case under the null hypothesis in this subsection.

For the lower bound, we will use $k < p$ to denote the number of summands which are still included. The lower bound is given by:

$$\sum_{s=0}^{p} \mathbb{P}(J_0 + J_1 > F_\alpha \mid |\widehat{S}| = s)\mathbb{P}(|\widehat{S}| = s) \geq \sum_{s=0}^{k} \mathbb{P}(J_0 + J_1 > F_\alpha \mid |\widehat{S}| = s)\mathbb{P}(|\widehat{S}| = s). \quad (10)$$

Since the later summands become negligible when $n$ gets larger, we expect this lower bound to be fairly "tight". Because we include only $k$ terms, this lower bound enables us to restrict our attention to only a few high-dimensional integrals.

Furthermore, an upper bound can also be easily obtained. For this upper bound, we use the fact that the conditional probabilities in (4) cannot be larger than 1. For this upper bound, we will use $l < p$ to denote the number of terms which we still compute. The upper bound is now given by:

$$\sum_{s=0}^{p} \mathbb{P}(J_0+J_1 > F_\alpha \mid |\widehat{S}| = s)\mathbb{P}(|\widehat{S}| = s) \leq \sum_{s=0}^{l} \mathbb{P}(J_0+J_1 > F_\alpha \mid |\widehat{S}| = s)\mathbb{P}(|\widehat{S}| = s)+ \sum_{s=l+1}^{p} \mathbb{P}(|\widehat{S}| = s). \tag{11}$$

Notice that when multiple components are included in $\widehat{S}$, it follows that $J_0$ becomes relatively large. Therefore, the conditional expectation of $\mathbb{P}(J_0 + J_1 > F_\alpha \mid |\widehat{S}| = s)$ becomes approximately equal to 1. Hence, this upper bound should also be fairly "tight".

By choosing a small value for $k$ and $l$, this can significantly reduce the computational cost. However, both the lower and upper bound are still computationally demanding. We have decreased the number of high-dimensional integrals but not the dimension of the integrals. Therefore, we will now take a closer look at the conditional probability of the upper and lower bound separately to see if we can obtain a bound in which the dimension of the integral is also reduced.

### 5.7.1 Upper bound

For the upper bound, we obtain the following:

$$\begin{aligned}
&\mathbb{P}(J_0 + J_1 > F_\alpha \mid |\widehat{S}| = s) \\
&= \mathbb{P}\left(\sqrt{p}n\sum_{j=1}^{s}\widehat{\theta}_j^2 + n\sum_{j=1}^{p}\widehat{\theta}_j^2 > F_\alpha \mid n\widehat{\theta}_1^2 > \delta_{p,n}^2, \ldots, n\widehat{\theta}_s^2 > \delta_{p,n}^2, n\widehat{\theta}_{s+1}^2 < \delta_{p,n}^2, \ldots, n\widehat{\theta}_p^2 < \delta_{p,n}^2\right) \\
&= \mathbb{P}\left((1+\sqrt{p})n\sum_{j=1}^{s}\widehat{\theta}_j^2 + n\sum_{j=s+1}^{p}\widehat{\theta}_j^2 > F_\alpha \mid n\widehat{\theta}_1^2 > \delta_{p,n}^2, \ldots, n\widehat{\theta}_s^2 > \delta_{p,n}^2, n\widehat{\theta}_{s+1}^2 < \delta_{p,n}^2, \ldots, n\widehat{\theta}_p^2 < \delta_{p,n}^2\right) \\
&\leq \mathbb{P}\left((1+\sqrt{p})n\sum_{j=1}^{s}\widehat{\theta}_j^2 + n\sum_{j=s+1}^{p}\widehat{\theta}_j^2 > F_\alpha \mid n\widehat{\theta}_1^2 > \delta_{p,n}^2, \ldots, n\widehat{\theta}_s^2 > \delta_{p,n}^2\right).
\end{aligned}$$

We used here that omitting the condition which tells us that some components are bounded from above by $\delta_{p,n}^2$, can only increase the probability of rejection. We can now rewrite the above expression to:

$$\mathbb{P}\left(n\sum_{j=s+1}^{p}\widehat{\theta}_j^2 > F_\alpha - (1+\sqrt{p})n\sum_{j=1}^{s}\widehat{\theta}_j^2 \mid n\widehat{\theta}_1^2 > \delta_{p,n}^2, \ldots, n\widehat{\theta}_s^2 > \delta_{p,n}^2\right)$$

On the left-hand side of the inequality in the probability, we have a sum of $\chi^2$ random variables. Since we now left out the condition about the magnitude of these $n\widehat{\theta}_j^2$ random variables with $j \in \{s+1, \ldots, p\}$, this left-hand size now just has a $\chi^2$ distribution with $p - s$ degrees of freedom. The following proposition uses this to obtain an upper bound which is also easily computable for higher values of $p$.

**Proposition 1.** *The conditional probability in the upper bound for the size is given by:*

$$\mathbb{P}\left(n\sum_{j=s+1}^{p}\widehat{\theta}_j^2 > F_\alpha - (1+\sqrt{p})n\sum_{j=1}^{s}\widehat{\theta}_j^2 \mid n\widehat{\theta}_1^2 > \delta_{p,n}^2,\ldots,n\widehat{\theta}_s^2 > \delta_{p,n}^2\right) =$$
$$\frac{1}{\overline{F}_{n\widehat{\theta}_1^2}(\delta_{p,n}^2)^s}\int_{\delta_{p,n}^2}^{\infty}\cdots\int_{\delta_{p,n}^2}^{\infty}\overline{F}_{\chi^2_{p-s}}\left(F_\alpha-(1+\sqrt{p})\sum_{j=0}^{s}x_j\right)\prod_{j=1}^{s}f_{n\widehat{\theta}_1^2}(x_j)dx_1\ldots dx_s.$$

In this upper bound, we reduced a $p$-dimensional integral to an $s$-dimensional integral. In the upper bound given in (11), we only need the conditional probability for small values of $s$. Therefore, this upper bound is also easily computable when we start increasing the value of $p$. We will now continue with the lower bound to see if we can achieve a similar result.

### 5.7.2 Lower bound

For the lower bound, we obtain the following:

$$\begin{aligned}
&\mathbb{P}(J_0+J_1>F_\alpha \mid |\widehat{S}|=s)\\
&=\mathbb{P}\left(\sqrt{p}n\sum_{j=1}^{s}\widehat{\theta}_j^2+n\sum_{j=1}^{p}\widehat{\theta}_j^2>F_\alpha \mid n\widehat{\theta}_1^2>\delta_{p,n}^2,\ldots,n\widehat{\theta}_s^2>\delta_{p,n}^2,n\widehat{\theta}_{s+1}^2<\delta_{p,n}^2,\ldots,n\widehat{\theta}_p^2<\delta_{p,n}^2\right)\\
&\overset{(1)}{\geq}\mathbb{P}\left((1+\sqrt{p})\sum_{j=1}^{s}\delta_{p,n}^2+n\sum_{j=s+1}^{p}\widehat{\theta}_j^2>F_\alpha \mid n\widehat{\theta}_{s+1}^2<\delta_{p,n}^2,\ldots,n\widehat{\theta}_p^2<\delta_{p,n}^2\right)\\
&\overset{(2)}{=}\mathbb{P}\left(n\sum_{j=s+1}^{p}\widehat{\theta}_j^2>C_\alpha \mid n\widehat{\theta}_{s+1}^2<\delta_{p,n}^2,\ldots,n\widehat{\theta}_p^2<\delta_{p,n}^2\right), \qquad (12)
\end{aligned}$$

where at (1) we replaced all $n\widehat{\theta}_j^2$ which are included in $\widehat{S}$ by $\delta_{p,n}^2$ and at (2) we introduced the constant $C_\alpha = F_\alpha - (1+\sqrt{p})s\delta_{p,n}^2$. By doing this, we reduced the $p$-dimensional integrals to $(p-s)$-dimensional integrals. However, the problem here is that the terms which we include in equation (10) are actually the terms in which only a few components are included in $\widehat{S}$, so these are the terms in which $s$ is small. Therefore, the $(p-s)$-dimensional integrals are still computationally demanding when we want to compute this for a large value of $p$.

Therefore, we will further rewrite this lower bound. In equation (12), we have to compute the probability that a sum is larger than a constant and we know that all the components from the sum are bounded by $\delta_{p,n}^2$. To get further intuition into this probability, it might be useful to represent it visually. In Figure 5.6, we consider the 2-dimensional case. Depending on the value of $C_\alpha$, there are four cases that we can separate.

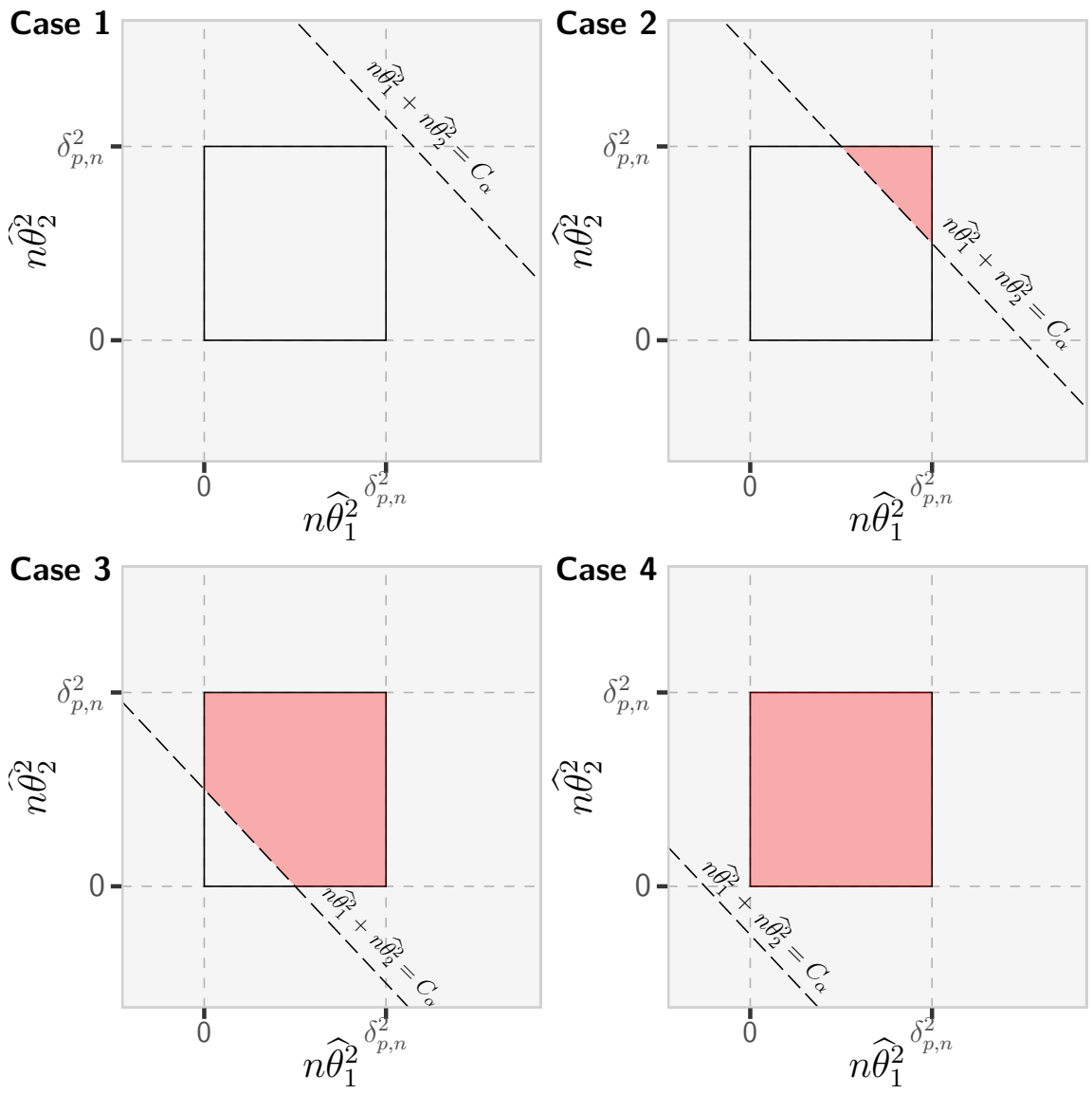


Figure 5.6: Four different cases where the pink area indicates the area we want to calculate conditional on being inside the box

Since we condition on being inside the box, it immediately follows that the probability is 0 for case 1 and that the probability is 1 for case 4. Furthermore, Proposition 2 tells us how to compute this probability for case 3.

**Proposition 2.** *The conditional probability of the lower bound in case 3, i.e., the case where* $0 < C_\alpha < \delta_{p,n}^2$*, is given by:*

$$\mathbb{P}\left( n \sum_{j=s+1}^{p} \widehat{\theta}_j^2 > C_\alpha \mid n\widehat{\theta}_{s+1}^2 < \delta_{p,n}^2, \ldots, n\widehat{\theta}_p^2 < \delta_{p,n}^2 \right) = 1 - \frac{F_{\chi_{p-s}^2}(C_\alpha)}{F_{\chi_1^2}\left(\delta_{p,n}^2\right)^{p-s}}.$$

Hence, we now know what the probability is when we are in case 1, 3, or 4 from Figure 5.6. Nevertheless, case 2 is also crucial to get a proper lower bound. This is because as $p$ gets larger, the term $s = 0$ starts to dominate as also shown in Figure 5.5. For this term,

the constant $C_\alpha$ is just equal to $F_\alpha$ and therefore it is large relative to the other terms. Consequently, the term $s = 0$ will mainly be in case 2. We will now take a closer look at this case.

In Figure 5.7, a more detailed illustration of case 2 can be found. This figure shows why it is not possible to use the result from Proposition 2 since we would then also take the blue areas into account.

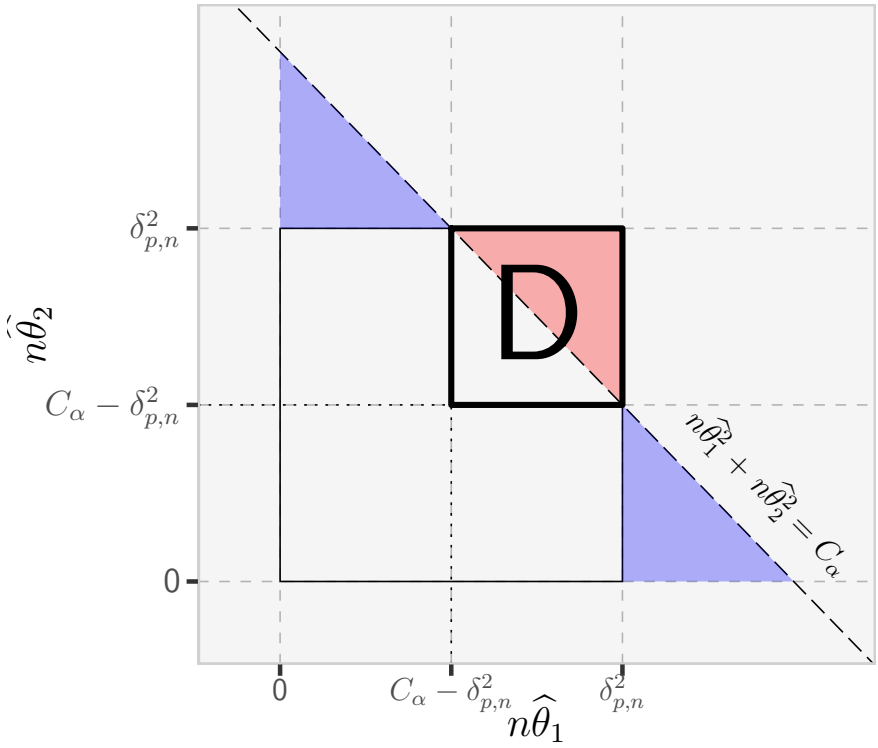


Figure 5.7: A closer look at case 2

We were unable to derive a closed-form expression for the probability in case 2. Therefore, we resort to the method of Monte Carlo to approximate this probability. Fortunately, there is also a trick to do this efficiently. For this trick, we use our knowledge about the distribution of all terms inside area D given in Figure 5.7. This is the rectangle bounded by the the outside of the box and the lines $n\widehat{\theta}_1^2 = C_\alpha - \delta^2_{p,n}$ and $n\widehat{\theta}_2^2 = C_\alpha - \delta^2_{p,n}$. For a general dimension $p$, this would be the area where each element $n\widehat{\theta}_j^2$ satisfies $C_\alpha - (p-1)\delta^2_{p,n} < n\widehat{\theta}_j^2 < \delta^2_{p,n}$. We can draw samples from this area by using that $n\widehat{\theta}_j^2$ follows a $\chi^2$ distribution with 1 degree of freedom and therefore, we have a joint truncated $\chi^2$ distribution inside area D. By drawing samples from this distribution, we can check whether the sample is in the pink area and we then obtain an approximation of the probability of being in the pink area conditional on

being in D. Using this, the conditional probability for the lower bound in case 2 is given by:

$$\begin{aligned}
&\mathbb{P}\left(n\sum_{j=s+1}^{p}\widehat{\theta}_j^2 > C_\alpha \mid n\widehat{\theta}_{s+1}^2 < \delta_{p,n}^2, \ldots, n\widehat{\theta}_p^2 < \delta_{p,n}^2\right) \\
&\quad = \frac{\mathbb{P}\left(n\sum_{j=s+1}^{p}\widehat{\theta}_j^2 > C_\alpha,\ n\widehat{\theta}_{s+1}^2 < \delta_{p,n}^2, \ldots, n\widehat{\theta}_p^2 < \delta_{p,n}^2\right)}{\mathbb{P}(n\widehat{\theta}_{s+1}^2 < \delta_{p,n}^2, \ldots, n\widehat{\theta}_p^2 < \delta_{p,n}^2)} \\
&\quad = \frac{\mathbb{P}\left(n\sum_{j=s+1}^{p}\widehat{\theta}_j^2 > C_\alpha,\ n\widehat{\theta}_{s+1}^2 < \delta_{p,n}^2, \ldots, n\widehat{\theta}_p^2 < \delta_{p,n}^2 \mid D\right)\mathbb{P}(D)}{\mathbb{P}(n\widehat{\theta}_{s+1}^2 < \delta_{p,n}^2, \ldots, n\widehat{\theta}_p^2 < \delta_{p,n}^2)},
\end{aligned}$$

where $\mathbb{P}\left(n\sum_{j=s+1}^{p}\widehat{\theta}_j^2 > C_\alpha,\ n\widehat{\theta}_{s+1}^2 < \delta_{p,n}^2, \ldots, n\widehat{\theta}_p^2 < \delta_{p,n}^2 \mid D\right)$ will be approximated using the Monte Carlo method described above. We will use this in the lower bound to compute the $s = 0$ term.

Now that we have derived an upper and lower bound, we want to examine their performance in practice. In Figure 5.8, these bounds are plotted with the Monte Carlo approximation for $p = 10$. For this upper bound we used $l = 2$, i.e., we included the first three terms. For the lower bound, we used $k = 10$ since the last terms are not harder to compute.

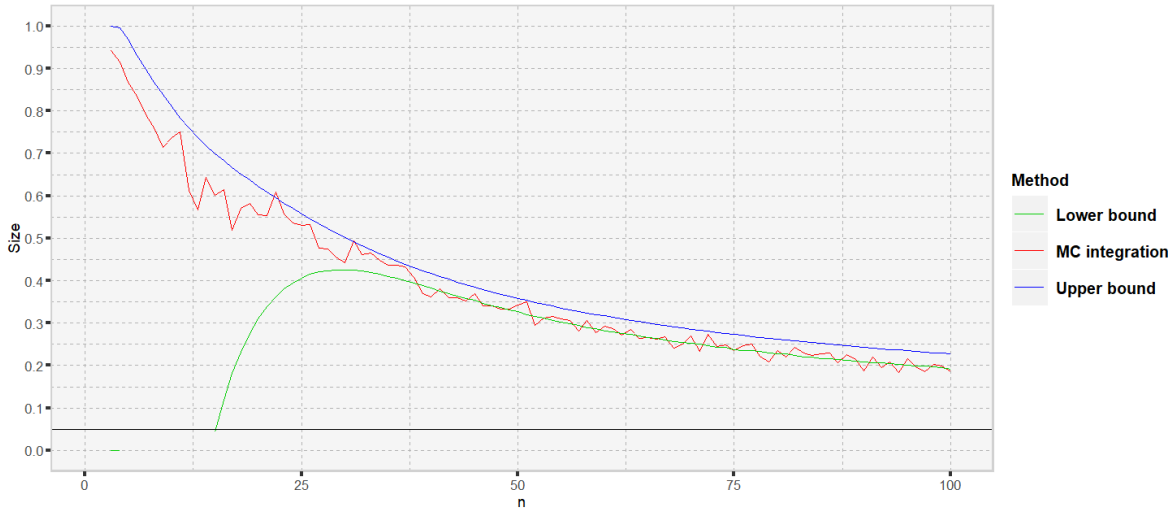


Figure 5.8: Upper and lower bound of the size for $p = 10$. For the bounds we used $l = 2$ and $k = 10$

As can be seen, both the upper and lower bound are pretty tight. The performance of the lower bound is not that well for small values of $n$. This is due to the fact that for most components, the value of $C_\alpha$ is relatively large and therefore, these components are still in case 2 at that point. When we use the result from Proposition 2 to compute these, the lower bound becomes quite loose. However, after some point, the lower bound becomes even tighter as the upper bound. This can be explained by the fact that we included all terms in the lower bound and only 3 terms in the upper bound.

Moreover, we observe that the variation in the lower bound due to the Monte Carlo approximation is very small. This confirms that our method of Monte Carlo approximation is indeed quite efficient. For this plot, we used 4,000 replications per value of $n$ and it took

about 4 minutes to generate the lower bound.

Hence, we now have useful bounds for the size which also work for higher values of $p$. Therefore, the bounds are also convenient for practitioners who want to gain insights into the size of their experiment.

## 5.8 Dependence on $p$ and $n$

In Figure 5.3, we saw that increasing $p$ has a different effect for small values of $n$ as for large values of $n$. In this subsection, we will obtain the intuition where this phenomenon comes from.

To examine how the probability of rejection depends on $p$ and $n$, it seems useful to take a closer look at the binomial distribution of $|\widehat{S}|$. If we would know how this probability distribution depends on $p$ and $n$, we would also know how $\mathbb{P}(J_0 = 0)$ depends on $p$ and $n$. This follows from the fact that $\mathbb{P}(J_0 = 0) = \mathbb{P}(|\widehat{S}| = 0)$.

To write the distribution of $|\widehat{S}|$ in terms of $p$ and $n$, we will first rewrite our earlier defined parameter $q = \mathbb{P}(\sqrt{n}|\widehat{\theta}_1| > \delta_{p,n})$. Under the null, we obtain that:

$$\begin{aligned} q &= \mathbb{P}(\sqrt{n}|\widehat{\theta}_1| > \delta_{p,n}) \\ &= 2\mathbb{P}(\sqrt{n}\widehat{\theta}_1 > \delta_{p,n}) \\ &= 2(1 - \Phi(\delta_{p,n})) \\ &= 2\left(1 - \frac{1}{2}\left(1 + \mathsf{erf}\left(\frac{\delta_{p,n}}{\sqrt{2}}\right)\right)\right) \\ &= 1 - \mathsf{erf}\left(\frac{\delta_{p,n}}{\sqrt{2}}\right). \end{aligned}$$

Now plugging this into the binomial distribution yields that:

$$\mathbb{P}(|\widehat{S}| = s) = \binom{p}{s}\left(1 - \mathsf{erf}\left(\frac{\delta_{p,n}}{\sqrt{2}}\right)\right)^s \mathsf{erf}\left(\frac{\delta_{p,n}}{\sqrt{2}}\right)^{p-s}.$$

Consequently, we obtain the probability that $J_0$ is zero as a function of $p$ and $n$:

$$\mathbb{P}(J_0 = 0) = \mathbb{P}(|\widehat{S}| = 0) = \left(\frac{2}{\sqrt{\pi}}\int_0^{\delta_{p,n}/\sqrt{2}} e^{-t^2} dt\right)^p. \tag{13}$$

Notice that the upper limit of the integral $\frac{\delta_{p,n}}{\sqrt{2}}$, is strictly positive. Therefore, the output of the error function will also be strictly positive. Furthermore, we know that the error function is bounded from above by 1. This tells us that we have some term which lies between 0 and 1 raised to the power $p$.

It can be easily verified that this probability is monotonically increasing in $n$. If $n$ gets larger, the domain of integration increases and for the rest, everything stays the same. This is also in line with the previous figures shown, as we keep $p$ constant and keep increasing $n$, the size distortion gets smaller and smaller. This is explained by the fact that when $n$ increases, the threshold $\delta_{p,n}$ also increases but the distribution of $\sqrt{n}|\widehat{\theta}_1|$ stays the same. Therefore, the probability that $\widehat{S}$ is empty also increases.

It gets interesting when we fix $n$ and start increasing $p$. When we do this, the domain of integration increases which gives a positive effect on the probability. Next to that, when

we increase $p$, we also raise expression (13) with a higher power. Since this expression lies between 0 and 1, there is also a decreasing effect. This provides some insights for what is happening in Figure 5.3. In this plot, increasing $p$ also increases the size distortion when $n$ is small but, it decreases the size distortion when $n$ is large. Hence, the value of $n$ determines what will happen when we start increasing $p$. And since $\delta_{p,n}$ determines the effect that a certain value of $n$ will have, this shows prominence of $\delta_{p,n}$. The threshold $\delta_{p,n}$ can be seen as an interaction term between $p$ and $n$.

This also makes sense when we look at equation (13). Consider the case where $n$ is relatively small. This means that we are increasing the domain of integration with a smaller factor as we would when we would use a larger value of $n$. In Figure 5.9, a plot of expression (13) is shown for three different values of $n$. In this plot, we first observe a quick rise of the probability when $p$ is small. This is because the error function is the integral over a bell curve and the probability mass is the highest around the zero point.

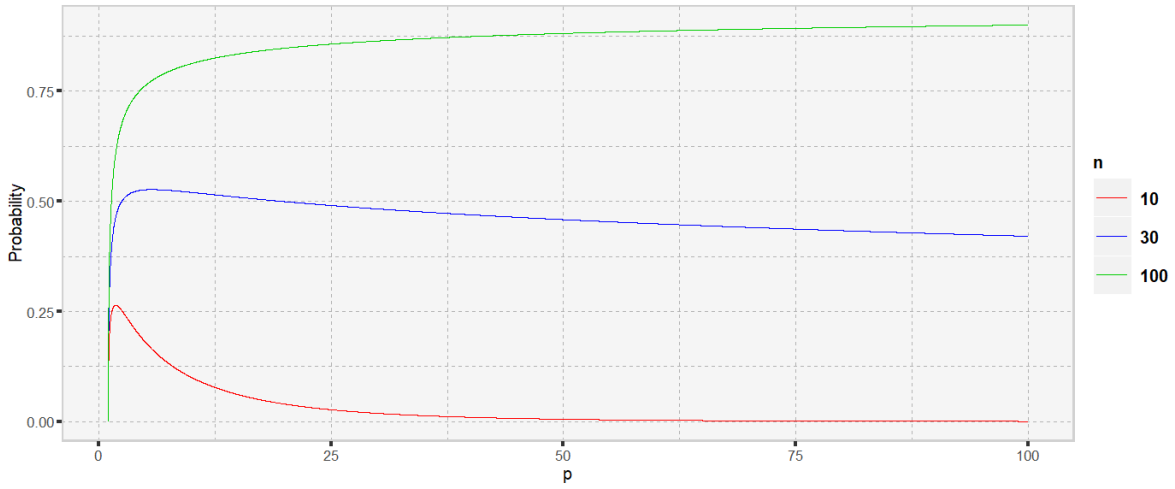


Figure 5.9: Probability of $J_0 = 0$ for different values of $n$

This figure shows some interesting results. In the case where $n = 10$, we observe that the probability goes to zero. This means that the exponent $p$ in (13) quickly starts to dominate the effect of the $p$ in the upper part of the integral. Notice that for a high value of $p$, the integral itself will be almost equal to 1. However, apparently, the effect of the exponent $p$ is dominant here and the whole term becomes quickly equal to 0. This also explains the large size in Figure 5.3 for small values of $n$. The size becomes almost equal to 1 in these plots because there will always be terms in $\widehat{S}$.

Furthermore, it follows from Figure 5.9 that it does not always have to be the case that the exponent $p$ is dominating. In fact, we observe that for $n = 100$, the probability keeps increasing as $p$ gets larger. In this case, we actually have that the $p$ in the upper part of the integral dominates the exponent $p$. Hence, there seems to be some kind of switch-point of $n$ from which the probability actually starts increasing with $p$ instead of decreasing.

This also gives us some more intuition about what is happening with our current choice of $\delta_{p,n}$. It seems that when $n$ is still small, $\delta_{p,n}$ does not grow fast enough with $p$ since the size converges to 1. So, the effect of adding more components, and therefore more components that can have an effect in terms of $J_0$, does not weigh against the growth of $\delta_{p,n}$ in $p$. For a higher value of $n$, this is different. Then $\delta_{p,n}$ does seem to be more

satisfactory since the size becomes smaller when we start increasing $p$. However, there is of course also a trade-off here between the power and the size. Therefore, we first need to consider the power for different values of $p$ and $n$ before we can draw any conclusions. This will be done in Section 5.9 and in the simulations of this study.

## 5.9 Analysis of power

With our current results, we can also analyze the power of the test. Note that we can only consider alternatives $\boldsymbol{\theta}$ which are dense since we need the assumption of identicality of components. We will, therefore, examine the power against the alternative:

$$\boldsymbol{\theta} = c\boldsymbol{\iota}_p,$$

for different values of scalar $c$. We can then compare this power with the power of the initial test. The power of the initial test can easily be computed since we also know the finite sample distribution under the alternative. This is given by the non-central $\chi^2$ distribution with $p$ degrees of freedom and non-centrality parameter $\sum_{j=1}^{p} n\theta_j^2 = pnc$. Since this distribution does not depend on $n$, the power is constant when we fix $p$ and start varying $n$. In Figure 5.10, the power and size of the power enhancement test and the initial test are shown. In this plot, the null corresponds to the $c = 0$ case and the alternative is chosen as the $c = 1$ case. Moreover, this plot corresponds to the case where $p = 2$ such that we can compute the exact size and power.

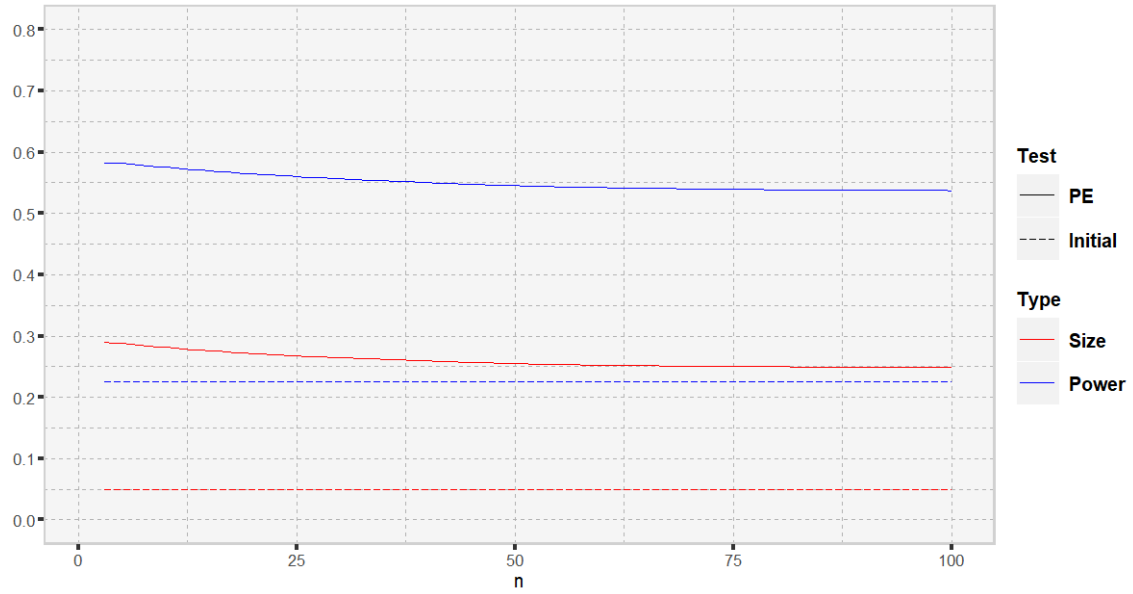


Figure 5.10: Power and size for the case $p = 2$

In this figure, we observe a large increase in both size and power. This is also due to the trade-off between size and power. Since we gave up quite some size, we earned a lot more power.

In this subsection, we only considered the dense alternative. However, in practice, it is often the case that the alternative is sparse instead of dense. Therefore, it is even more interesting to look at the increase in power for the sparse alternative. Since we cannot analyze these alternatives with our current results, we will resort to simulations.

# 6 Simulations

In this section, we consider multiple settings in which the power enhancement technique will be investigated. In the previous subsection, we had to make some simplifying assumptions to be able to analyze the size distortion theoretically. In our simulations, we start with this basic setting to numerically verify the theoretical results. After that, we investigate the power against sparse alternatives in this setting. Finally, we scrutinize how sensitive the technique is to the choice of the threshold $\delta_{p,n}$.

## 6.1 Verification of results

This setting is the same as the setting in Section 5 with the corresponding assumptions of independence and identicality and normality. Therefore, it is interesting to compare these results with the theoretical results. We have simulated the case where $p = 10$ for the value of $n$ ranging from 1 to 100. Furthermore, we used 2,000 replications per simulation. In Figure 6.1, we have combined the results from the simulations with the derived bounds and the method of Monte Carlo integration.

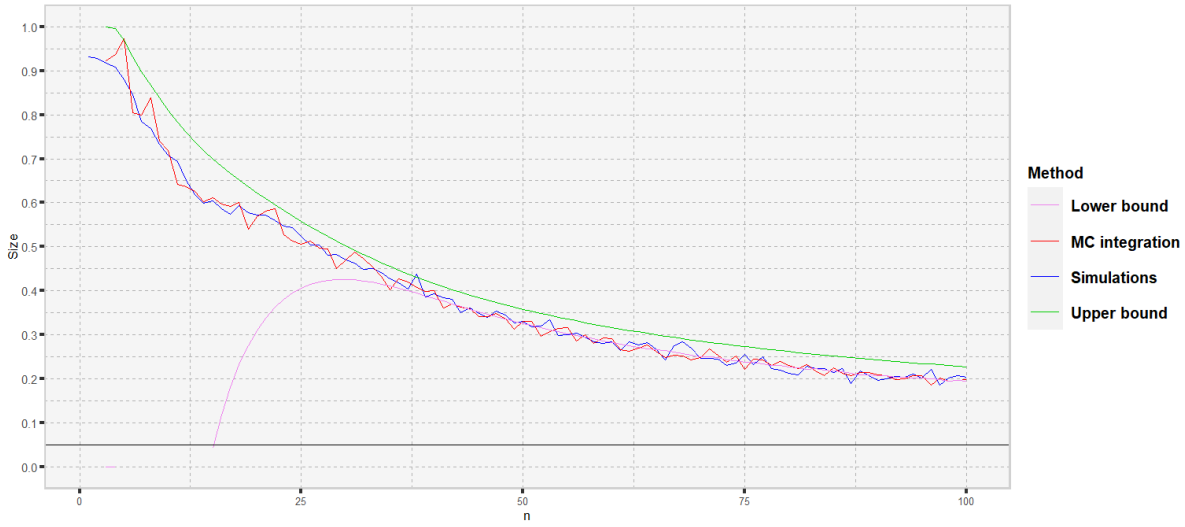


Figure 6.1: Comparison of the size of the PE technique between several methods, for $p = 10$. For the bounds we used $l = 2$ and $k = 10$

In this figure, we see that the simulations line up nicely with the approximations using Monte Carlo integrations. For a small value of $n$, there seems to be a higher variance in the approximations using Monte Carlo integrations. Furthermore, we used only 2,000 replications for the simulations and 30,000 for the Monte Carlo integration. Because of this computational advantage, we will not use the method of Monte Carlo integration anymore for the rest of this section.

Furthermore, the upper and lower bound also lines up very well with the simulation results. In general, there seems to be a gap of about 0.02 for the upper bound and the gap with the lower bound becomes negligible after some point. The advantage of these bounds is that they are computationally pleasant.

Nevertheless, we do not know how tight these bounds are for higher values of $p$. Therefore, we will also compare it with the simulation results for $p = 200$. The result can be found in Figure 6.2. To generate this plot, we again used 2,000 replications. Also, note that $n$ goes to 250 here.

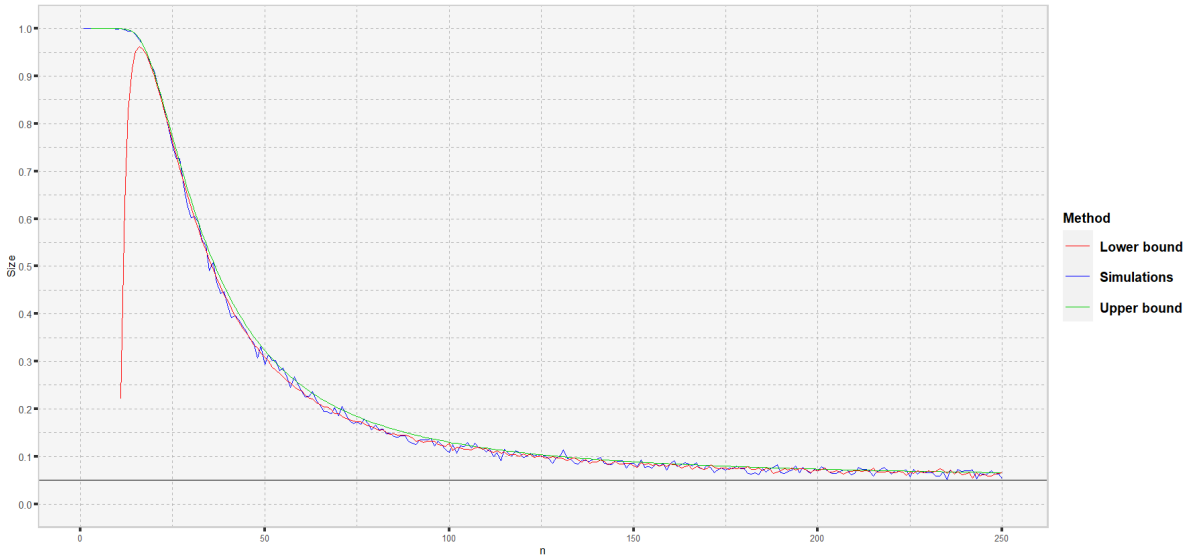


Figure 6.2: Verification of the bounds for $p = 200$. For the bounds we used $l = 2$ and $k = 10$

We observe that both bounds are very tight. Computationally, this is a very pleasing result. The simulations took about 3.5 hours on a laptop with an Intel Core i7 processor. In comparison, the computation of the upper bound only took 5 seconds and the lower bound only 13 minutes. The difference between these mainly comes from the Monte Carlo method used for the lower bound. Because of the fast computation of these bounds, they provide a very elegant tool for the practitioner to keep track of the size in their experiments.

In Figure 6.2, we also observe that mainly the upper bound is much tighter as in the case with $p = 10$. This provides us with some more insight. In the derivation of the upper bound, we find two contrasting effects as $p$ increases. On the one hand, we neglect $p - s$ conditions. So, as $p$ increases, this has a negative effect on the tightness of the upper bound. However, these conditions are given by $\widehat{\theta}_j^2 < \delta_{p,n}^2$ and as $p$ increases, the right-hand side of this inequality also increases and the left-hand side stays the same. In that sense, this condition also becomes weaker. From Figure 6.2, it seems like the latter effect is the dominating effect.

In the lower bound, we only made two inequality steps. The first one was omitting the later summands but we saw that this only affects the tightness of the bound for very small values of $n$. The second inequality step was replacing the $n\widehat{\theta}_j^2$ terms that satisfy $n\widehat{\theta}_j^2 > \delta_{p,n}^2$, i.e. the terms that are included in $\widehat{S}$, by $\delta_{p,n}^2$. For large values of $p$ and $n$, the term with 0 components in $\widehat{S}$ ($s = 0$) and the term with 1 component in $\widehat{S}$ ($s = 1$), are dominating. For the $s = 0$ term, the second inequality step holds with equality as no components are replaced by $\delta_{p,n}^2$. For the $s = 1$ term, only component has to be replaced so this will also not make a large difference. Therefore, it makes sense that the lower bound is very tight for large values of $p$ and $n$.

## 6.2 Different sparsity levels

In the theoretical analysis, we were only able to analyze alternatives in which all elements are equal to each other. We will now investigate the performance of the technique against sparse alternatives. In this investigation, we still consider the same data generating process as before and we also assume normality.

To compare alternatives in terms of sparsity, the $\ell_0$-norm will be used. We will only be comparing vectors consisting of zeros and non-zero elements of comparable absolute magnitude. As explained in Section 2.1.1, the $\ell_0$-norm is an appropriate sparsity measure in this setting. The non-zero elements in this vector cannot be too large since this would lead to tables filled with 100% rejection rates. By trial and error, we found that vectors consisting of zeros and 0.05 satisfy this criterion and have rejection rates that are not too small neither too large.

In this simulation, we will consider the values $p = 100, 300$, and $500$ and $n = 150, 250$, and $400$. We have chosen $n$ to be at least $150$ such that the size stays manageable, as can be seen in Figure 6.2. By letting both of $p$ and $n$ vary, we can study how the size and power depend on these variables.

We start by generating data under the null. It is essential to do this such that we are also able to compare the power and the size. Afterwards, we generate data under three different alternative hypotheses. These alternatives differ from each other in terms of sparsity. To compare these vectors, we will use $\mathcal{D}_s$ to denote the space of vectors with $s\%$ of the components equal to $0.05$. In this vector, we will set the first components equal to $0.05$ and the last components equal to $0$. The alternatives are defined by:

$$
\begin{aligned}
H_1^a &: \boldsymbol{\theta} \in \mathcal{D}_5, \\
H_1^b &: \boldsymbol{\theta} \in \mathcal{D}_{15}, \\
H_1^c &: \boldsymbol{\theta} \in \mathcal{D}_{30}.
\end{aligned}
$$

Notice that these alternatives are ranked from most sparse to most dense. For the null and for these alternatives, we computed the percentage of rejections of using only the Wald test and the percentage of rejections when adding $J_0$ to the Wald test. The results of these simulations can be found in Table 1. For each value of $p$ and $n$, we used 2,000 replications.

| | | $H_0$ | | $H_1^a$ | | $H_1^b$ | | $H_1^c$ | |
|---|---|---|---|---|---|---|---|---|---|
| $p$ | $n$ | Wald | PE | Wald | PE | Wald | PE | Wald | PE |
| 100 | 150 | 5.3% | 9.1% | 6.3% | 11.8% | 10.4% | 16.3 % | 21.1 % | 26.4 % |
| | 250 | 4.8% | 6.7% | 7.7% | 9.5 % | 17.0 % | 20.4 % | 34.7 % | 37.0 % |
| | 400 | 5.5% | 6.6% | 9.8 % | 11.1 % | 28.3 % | 29.6 % | 62.0 % | 62.5 % |
| 300 | 150 | 5.8% | 8.7% | 7.5 % | 10.9 % | 19.1 % | 22.2 % | 39.7 % | 41.9 % |
| | 250 | 5.0% | 6.4% | 10.6 % | 12.4 % | 30.6 % | 32.1 % | 69.2 % | 69.9 % |
| | 400 | 5.1% | 5.9% | 15.2 % | 16.1 % | 55.9 % | 56.2 % | 94.1 % | 94.2 % |
| 500 | 150 | 5.3% | 8.0% | 8.7 % | 12.1 % | 22.5 % | 25.5 % | 52.1 % | 54.0 % |
| | 250 | 4.9% | 6.3% | 10.8 % | 11.8 % | 40.3 % | 41.5 % | 86.4 % | 86.6 % |
| | 400 | 5.2% | 5.7% | 17.5 % | 18.2 % | 73.2 % | 73.4 % | 99.3 % | 99.4 % |

Table 1: Size and power for different sparsity levels

We observe that the power enhancement test never has a lower percentage of rejections than the Wald test. This is of course how it should be. Furthermore, also notice that as we move more to the right in the table, we look at alternatives which include more $0.05$ terms and therefore it gives a higher percentage of rejections. Moreover, the results under the null also seem to be in line with the theoretical results. From Section 5.8, we know that the size should be decreasing in $p$ for these values of $n$. This is indeed the case. Next to that, we again see that the size distortion is the largest for the case where $n$ is small. In these cases, we also observe the largest enhancements in power.

Now, consider the alternative hypotheses, here we see that the power of the Wald test is increasing in $n$. This is as expected. However, for the power enhancement test, there is an interesting effect. For the hypothesis $H_1^a$, the hypothesis which is the closest to the null, we see that the power can also decrease with $n$ (see e.g. $p = 100$ and $n = 150, 250$). This can be explained in the following way. It is known that a component $j$ is included in $\widehat{S}$ if $n\widehat{\theta}_j^2 > \delta_{p,n}^2$. This left-hand side has the same distribution for every value of $n$. However, the right-hand side increases with $n$ and therefore the inclusion probability becomes smaller and smaller. Nevertheless, it is not the case that the power is monotonically decreasing with $n$. For most values in Table 1, the power actually grows with $n$. This is still in line with our reasoning above. This growth in $n$ is due to the growth in power of the Wald test. The observation that $\mathbb{P}(n\widehat{\theta}_j^2 > \delta_{p,n}^2)$ is decreasing with $n$ only means that the difference between $J_0$ and $J_0 + J_1$ becomes smaller.

At the moment, we are not able yet to fairly compare the performance of the technique for different sparsity levels. This is because the alternatives used in Table 1 have different $\ell_2$-norms. Therefore, the power of the Wald test also differs between those alternatives. It does seem like the difference between the Wald and the PE technique is slightly larger for the sparser alternatives. However, since the power of the Wald test also differs substantially here, we cannot just linearly compare these differences. In the next subsection, we will address this problem and combine this with examining the performance of the technique for different thresholds $\delta_{p,n}$.

## 6.3 Different threshold sequences $\delta_{p,n}$

As we have seen in Section 4.1, the size and power seem to be sensitive to the choice of $\delta_{p,n}$. In their paper, Fan et al. (2015) only give asymptotic arguments for this choice of threshold. However, there exist a multitude of threshold parameters that satisfy the same asymptotic properties. For example, $c\delta$, for any fixed $c > 0$. In the simulations of their study, Fan et al. (2015) also used these different thresholds with $c = 0.9, 1, 1.06$, and $\sqrt{1.5}$. This is worrying for practitioners, as it is impossible to check which threshold value is appropriate for their setting. Therefore, we will examine how sensitive this test is to the threshold.

Just as Fan et al. (2015), we will perform our simulations using the values $c = 0.9, 1, 1.06$, and $\sqrt{1.5}$. We will consider this for the values $p = 200, 500$ and $n = 150, 250, 400$. Moreover, we will examine this setting for three different underlying values of $\boldsymbol{\theta}$. The first one will be the null, $\boldsymbol{\theta} = \mathbf{0}$. The second will be a sparse alternative and the third one will be a rather dense alternative. They will be generated under the following two hypotheses:

$$H_1^s : \boldsymbol{\theta} \in \mathcal{D}_1,$$
$$H_1^d : \boldsymbol{\theta} \in \mathcal{D}_{25}.$$

For the sparse alternative, we will use a vector with first $0.2$ terms and afterwards zeros. We have chosen higher values as before to make sure that the power of the Wald test does not become too small. For the dense alternative, we will use something else. In Table 1, the power of the dense alternatives was always higher since these vectors had a higher $\ell_2$-norm. To be able to fairly compare the dense and the sparse vector, we will now make the adjustment by using a dense vector with the same $\ell_2$-norm. The power of the Wald test then stays constant since it only depends on the Euclidean length of a vector. If we still want a constant $c$ and zeros in this vector, we need to solve the following equation:

$$\sqrt{\frac{p}{100}\left(0.2\right)^2} = \sqrt{\frac{p}{4}c^2} \iff \frac{p}{2500} = \frac{p}{4}c^2 \iff c = 0.04.$$

Therefore, the vectors under the dense alternative will consist of 0.04 terms and afterwards zeros. In Table 2, the results can be found. In this table, the subscript of PE denotes the constant $c$ in front of the threshold.

| | | Wald | $\text{PE}_{\sqrt{1.5}}$ | $\text{PE}_{1.06}$ | $\text{PE}_{1}$ | $\text{PE}_{0.9}$ |
|---|---|---|---|---|---|---|
| $p$ | $n$ | | | $H_0$ | | |
| 200 | 150 | 4.8% | 4.8 % | 5.8 % | 8.2 % | 18.0 % |
| | 250 | 5.4 % | 5.5 % | 6.1 % | 7.1 % | 11.8 % |
| | 400 | 4.7 % | 4.7 % | 4.8 % | 5.1 % | 7.6 % |
| 500 | 150 | 5.1 % | 5.2 % | 6.4 % | 8.1 % | 18.1 % |
| | 250 | 4.7 % | 4.7 % | 5.1 % | 5.8 % | 10.0 % |
| | 400 | 5.5 % | 5.5 % | 5.7 % | 5.9 % | 8.3 % |
| | | | | $H_1^s$ | | |
| 200 | 150 | 15.7 % | 17.9 % | 26.7 % | 32.7 % | 49.6 % |
| | 250 | 24.9 % | 29.8 % | 42.6 % | 51.0 % | 66.9 % |
| | 400 | 44.8 % | 52.9 % | 70.4 % | 77.9 % | 88.2 % |
| 500 | 150 | 22.1 % | 24.3 % | 33.7 % | 40.7 % | 61.9 % |
| | 250 | 45.6 % | 48.1 % | 60.4 % | 68.1 % | 82.7 % |
| | 400 | 76.3 % | 80.6 % | 90.4 % | 94.1 % | 98.4 % |
| | | | | $H_1^d$ | | |
| 200 | 150 | 16.0 % | 16.0 % | 17.9 % | 20.4 % | 31.3 % |
| | 250 | 24.5% | 24.5 % | 25.3 % | 26.5 % | 32.4 % |
| | 400 | 47.0 % | 47.0 % | 47.4 % | 47.7 % | 50.1 % |
| 500 | 150 | 23.2 % | 23.2 % | 24.0 % | 25.9 % | 36.7 % |
| | 250 | 45.7 % | 45.7 % | 45.9 % | 46.6 % | 51.1 % |
| | 400 | 76.7 % | 76.7 % | 76.7 % | 76.8 % | 77.8 % |

Table 2: Size and power for different thresholds

In this table, we observe the same patterns with $p$ and $n$ as in Table 1. Moreover, we observe that when the threshold decreases, the percentage of rejections increases. This is as expected. It immediately strikes our attention that the differences in rejection rates are very large for the sparse alternative. In their paper, Fan et al. (2015) both used the $0.9\delta_{p,n}$ threshold and the $\sqrt{1.5}\delta_{p,n}$. However, this can make a difference of almost 40% in terms of power. Yet, we also observe that these differences are rather small for the dense alternative. Next to that, it also seems like the sensitivity to the threshold decreases as $p$ and $n$ become larger. This makes sense since all thresholds have the same asymptotic properties. Hence, this table shows that especially for a quite small sample, the technique is very sensitive to the choice of $\delta_{p,n}$. For example, when $p = 200$ and $n = 150$, choosing the wrong threshold can lead to some substantial size distortion. This size distortion will even be larger when

$p$ and $n$ are also smaller. Because of this, the practitioner should be very careful with this threshold.

Furthermore, we can also compare the sparse alternative with the dense alternative. We clearly see that the sparse alternative is rejected more often by the power enhancement test. This table shows that the technique can be very useful when we are testing against a sparse alternative. Especially, when $p$ and $n$ are large, a small sacrifice in terms of size can boost the power a lot. For example, when $p = 500$ and $n = 250$, using the PE technique with $c = 1$ instead of the Wald test increases the size by 5.8% - 4.7% $= 1.1\%$. However, it also gives a boost in power of 68.1 % - 45.6 % $= 22.5$ %. This is definitely an interesting result and the practitioner might willingly give up some size here. For the dense alternative, the technique barely increases the power of the test and therefore the technique is quite futile in this setting.

We conclude that the technique is very sensitive to the threshold when $p$ and $n$ are rather small or when the alternative is very sparse. Moreover, the technique is especially useful when $p$ and $n$ are quite large and we are testing against a sparse alternative.

# 7 Conclusion

In this paper, we analyzed the power enhancement technique introduced by Fan et al. (2015) and provided some background. The power enhancement technique seems to be particularly helpful in a high-dimensional setting to boost the power against sparse alternatives. By the results of (Kock and Preinerstorfer, 2019), asymptotically, this technique can enhance the power of a lot of existing tests without raising the size. Nevertheless, it remained unclear how these properties carry over to finite samples. Therefore, this paper studied the performance of this technique in samples of a practical size.

In the analysis of this paper, we zoomed in on a special case. Under the assumptions that all components are equal and that the errors are i.i.d. standard normally distributed, a closed-form expression for the size and power had been derived. However, this expression becomes computationally very demanding when $p$ starts to increase. Therefore, we used the method of Monte Carlo integration to approximate it. Furthermore, we derived an upper and lower bound which are computationally pleasant, even when $p$ increases. We also provided intuition on how the performance of the technique depends on the number of parameters $p$ and the number of observations $n$. Further research can consider relaxing some of our assumptions to test the robustness of our results.

In this study, we found that there is some serious size distortion for small values of $p$ and $n$. This size distortion goes hand in hand with a large rise in power. Therefore, the practitioner should be aware that this power does not come for 'free' in practice and, that he might has to make a substantial sacrifice in terms of size. However, this all depends on the setting of the practitioner. In this paper, we also found that the size stays manageable for higher dimensions of $p$ and $n$ and that the technique can be very useful here.

After the theoretical analysis, we also performed several Monte Carlo simulations. With these simulations, we showed that the bounds are very tight for large values of $p$. This provides practitioners with an elegant tool to control the size in their experiments.

Subsequently, we used Monte Carlo simulations to further investigate the power of the technique under different sparsity levels and different choices for the threshold sequence $\delta_{p,n}$. It was shown that the sparsity of the alternative made a vast difference for the increase in

power. Moreover, we saw that the technique is very sensitive to the choice of the threshold and the practitioner should be cautious with this choice. Further research can investigate this threshold more in-depth and possibly find a threshold that improves the performance of the technique.

To conclude it all, we can say that the technique may not be that promising in practice as it seems asymptotically. When $p$ and $n$ are still relatively small, the practitioner should be aware that he has to make substantial sacrifices in terms of size. This goes hand in hand with the threshold sequence $\delta_{p,n}$, the technique is very sensitive to this choice and the practitioner should handle this with care. However, in a high-dimensional setting, the size stays manageable and the technique can be very useful to boost the power against sparse alternatives.

# References


Bai, Z. and H. Saranada (1996). Effect of high dimension: by an example of a two sample problem. *Statistica Sinica 6*, 311–329.

Cai, T., W. Liu, and Y. Xia (2014). Two-sample test of high dimensional means under dependence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 76*, 349–372.

Cai, T., W.Liu, and Y.Xia (2013). Two-sample covariance matrix testing and support recovery in high-dimensional and sparse settings. *Journal of the American Statistical Association 108*, 265–277.

Castagna, John P., Shengjie Sun, and Robert W. Siegfried (2003). Instantaneous spectral analysis: Detection of low-frequency shadows associated with hydrocarbons. *The leading edge 22*, 120–127.

Chakraborty, A. and P. Chaudhuri (2017). Tests for high-dimensional data based on means, spatial signs and spatial ranks. *Annals of Statistics 45*, 771–799.

Chen, Y. and Guo Y. (2019). Independence test in high-dimension using distance correlation and power enhancement technique. *Communications in Statistics - Theory and Methods 0*(0), 1–18.

Chernozhukov, Victor, Denis Chetverikov, and Kengo Kato (2013). Testing many moment inequalities. CeMMAP working papers CWP65/13, Centre for Microdata Methods and Practice, Institute for Fiscal Studies.

Dalton, Hugh (1920, 09). The Measurement of the Inequality of Incomes. *The Economic Journal 30*(119), 348–361.

Dempster, A.P. (1958). A high dimensional two sample significance test. *Annals of Mathematical Statistics 29*, 995–1010.

Fama, E.F. and K.R. French (1992). The cross-section of expected stock returns. *The Journal of Finance 47 (2)*, 427–465.

Fan, J. (1996). Test of significance based on wavelet thresholding and neyman's truncation. *Journal of the American Statistical Association 91*(434), 674–688.

Fan, J., Y. Liao, and J. Yao (2014). Power enhancement in high-dimensional cross-sectional tests. Technical report, Princeton University.

Fan, J., Y. Liao, and J. Yao (2015). Power enhancement in high-dimensional cross-sectional tests. *Econometrica 83*, 1497–1541.

Fan, Yingying and Cheng Yong Tang (2012, Dec). Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 75*(3), 531–552.

Hansen, Peter Reinhard (2003, 02). Asymptotic tests of composite hypotheses. *SSRN Electronic Journal*.

Hansen, Peter Reinhard (2005). A test for superior predictive ability. *Journal of Business & Economic Statistics 23*(4), 365–380.

Hurley, Niall P. and Scott T. Rickard (2008). Comparing measures of sparsity.

James, D., B.D. Clymer, and P. Schmalbrock (2001). Texture detection of simulated microcalcification susceptibility effects in magnetic resonance imaging of breasts. *Journal of Magnetic Resonance Imaging 13*, 876–881.

Juodis, A. and S. Reese (2018). The Incidental Parameters Problem in Testing for Remaining Cross-section Correlation. *arXiv e-prints*, arXiv:1810.03715.

Kock, A.B. and D. Preinerstorfer (2019). Power enhancement in high-dimensional cross-sectional tests. *Econometrica 87*, 1055–1069.

Koning, Nick (2019). Directing power towards conic parameter subspaces.

Liu, Y., W. Sun, P. R. Alexander, C.L. Kooperberg, and Q. He (2019). Statistical inference of genetic pathway analysis in high dimensions. *Biometrika 106 3*, 651.

Srivastava, M. S. and M. Du (2008). A test for the mean vector with fewer observations than the dimension. *Journal of Multivariate Analysis 99*, 386–402.

Steinberger, L. (2016). The relative effects of dimensionality and multiplicity of hypotheses on the f-test in linear regression. *Electronic Journal of Statistics 10*, 2584–2640.

Su, J., Y. Zhang, and J. Wei (2016). A practical test for strict exogeneity in linear panel data models with fixed effects. *Economics Letters 147*, 27–31.

Wang, Yishu, Dejie Yang, and Minghua Deng (2015, 08). Low-rank and sparse matrix decomposition for genetic interaction data. *BioMed research international 2015*, 573956.

Xu, Gongjun, Lifeng Lin, Peng Wei, and Wei Pan (2017, 03). An adaptive two-sample test for high-dimensional means. *Biometrika 103*(3), 609–624.

Yin, Penghang, Ernie Esser, and Jack Xin (2014, 01). Ratio and difference of l1 and l2 norms and sparse representation with coherent dictionaries. *Communications in Information and Systems 2*, 87–109.

Zhang, G-Ming, Shuyi Zhang, and Yuwen Wang (2000, 12). Application of adaptive time-frequency decomposition in ultrasonic nde of highly-scattering materials. *Ultrasonics 38*, 961–4.

Zhong, P., S. Chen, and M. Xu (2013). Tests alternative to higher criticism for high-dimensional means under sparsity and column-wise dependence. *The Annals of Statistics 41*, 2820–2851.

Zhong, P.S. and S. X. Chen (2011). Tests for high-dimensional regression coefficients with factorial designs. *Journal of the American Statistical Association 106*, 260–274.

# 8 Appendix

*Proof of Lemma 1.* This result trivially holds for $s = 0$ and $s = p$ since there is only one possible combination in which these can occur. That is, all components are not included in $\widehat{S}$ or all components are included in $\widehat{S}$. Now consider the case where $s \in \{1, \ldots, p-1\}$. Note that the number of different combinations in which these can occur is given by $\binom{p}{s} = C(p, s)$. We will denote the event of one such combination by $A_j$ with $j \in \{1, \ldots, C(p, s)\}$. Here, $j = 1$ corresponds to the case of the first $s$ elements included in $\widehat{S}$. Furthermore, note that these events are mutually exclusive. Using this notation, it follows that the event $|\widehat{S}| = s$ is equivalent to the union of the events of combinations. This means that:

$$|\widehat{S}| = s \iff \bigcup_{j=1}^{C(p,s)} A_j$$

For the conditional probability, we know that:

$$\mathbb{P}(J > F_\alpha \mid |\widehat{S}| = s) = \frac{\mathbb{P}(J > F_\alpha \cap |\widehat{S}| = s)}{\mathbb{P}(|\widehat{S}| = s)}$$

The denominator of this expression easily follows:

$$\mathbb{P}(|\widehat{S}| = s) = \mathbb{P}\left(\bigcup_{j=1}^{C(p,s)} A_j\right) \overset{\text{mut.excl.}}{=} \sum_{j=1}^{C(p,s)} \mathbb{P}(A_j) \overset{\text{ident.}}{=} C(p, s)\mathbb{P}(A_1)$$

For the numerator, we find that:

$$\begin{aligned}\mathbb{P}(J > F_\alpha \cap |\widehat{S}| = s) = \mathbb{P}\left(J > F_\alpha \cap \left(\bigcup_{j=1}^{C(p,s)} A_j\right)\right) &\overset{\text{set theory}}{=} \mathbb{P}\left(\bigcup_{j=1}^{C(p,s)} (J > F_\alpha \cap A_j)\right) \\ &\overset{\text{mut.excl.}}{=} \sum_{j=1}^{C(p,s)} \mathbb{P}(J > F_\alpha \cap A_j) \\ &\overset{\text{ident.}}{=} C(p, s)\mathbb{P}(J > F_\alpha \cap A_1)\end{aligned}$$

By the above results, we conclude that the conditional probability is equal to:

$$\begin{aligned}\frac{\mathbb{P}(J > F_\alpha \cap |\widehat{S}| = s)}{\mathbb{P}(|\widehat{S}| = s)} &= \frac{C(p,s)\mathbb{P}(J > F_\alpha \cap A_1)}{C(p,s)\mathbb{P}(A_1)} \\ &= \frac{\mathbb{P}(J > F_\alpha \cap A_1)}{\mathbb{P}(A_1)} \\ &= \mathbb{P}(J > F_\alpha \mid n\widehat{\theta}_1^2 > \delta_{p,n}^2, \ldots, n\widehat{\theta}_s^2 > \delta_{p,n}^2, n\widehat{\theta}_{s+1}^2 < \delta_{p,n}^2, \ldots, n\widehat{\theta}_p^2 < \delta_{p,n}^2)\end{aligned}$$

□

*Proof of Proposition 1.* We will start of by rewriting the conditional probability for the sake of notational simplicity:

$$\begin{aligned}&\mathbb{P}\left(n\sum_{j=s+1}^{p}\widehat{\theta}_j^2 > F_\alpha - (1+\sqrt{p})n\sum_{j=1}^{s}\widehat{\theta}_j^2 \mid n\widehat{\theta}_1^2 > \delta_{p,n}^2, \ldots, n\widehat{\theta}_s^2 > \delta_{p,n}^2\right) \\ &= \mathbb{P}\left(n\sum_{j=s+1}^{p}\widehat{\theta}_j^2 > F_\alpha - (1+\sqrt{p})n\sum_{j=1}^{s}\widehat{\theta}_j^2 \mid \overline{C}_s\right),\end{aligned}$$

where we introduced $\overline{C}_s = n\widehat{\theta}_1^2 > \delta_{p,n}^2, \ldots, n\widehat{\theta}_s^2 > \delta_{p,n}^2$. The left-hand side of the inequality in the probability has a $\chi^2$ distribution with $p-s$ degrees of freedom. Before we can use this, we first need to rewrite the expression a bit more. Using the third axiom of probability we know that:

$$\begin{aligned}&\mathbb{P}\left(n\sum_{j=s+1}^{p}\widehat{\theta}_j^2 > F_\alpha - (1+\sqrt{p})n\sum_{j=1}^{s}\widehat{\theta}_j^2 \mid \overline{C}_s\right) \\ &= \int_0^\infty \ldots \int_0^\infty \mathbb{P}\left(n\sum_{j=s+1}^{p}\widehat{\theta}_j^2 > F_\alpha - (1+\sqrt{p})n\sum_{j=1}^{s}\widehat{\theta}_j^2,\ n\widehat{\theta}_1^2 = x_1, \ldots, n\widehat{\theta}_s^2 = x_s \mid \overline{C}_s\right) dx_1 \ldots dx_s \\ &= \int_0^\infty \ldots \int_0^\infty \mathbb{P}\left(n\sum_{j=s+1}^{p}\widehat{\theta}_j^2 > F_\alpha - (1+\sqrt{p})\sum_{j=1}^{s}x_j,\ n\widehat{\theta}_1^2 = x_1, \ldots, n\widehat{\theta}_s^2 = x_s \mid \overline{C}_s\right) dx_1 \ldots dx_s \\ &= \frac{1}{\mathbb{P}(\overline{C}_s)}\int_0^\infty \ldots \int_0^\infty \mathbb{P}\left(n\sum_{j=s+1}^{p}\widehat{\theta}_j^2 > F_\alpha - (1+\sqrt{p})\sum_{j=1}^{s}x_j,\ n\widehat{\theta}_1^2 = x_1, \ldots, n\widehat{\theta}_s^2 = x_s, \overline{C}_s\right) dx_1 \ldots dx_s,\end{aligned}$$

where we used Bayes' Theorem in the last step. Now, recall that $\overline{C}_s$ is used to replace the conditions $n\widehat{\theta}_1^2 > \delta_{p,n}^2, \ldots, n\widehat{\theta}_s^2 > \delta_{p,n}^2$. Because of this, the above integral is zero for $(x_1, \ldots, x_s) \notin (\delta_{p,n}^2, \infty)^s$. Therefore, we can rewrite the above expression to:

$$\frac{1}{\mathbb{P}(\overline{C}_s)}\int_{\delta_{p,n}^2}^\infty \ldots \int_{\delta_{p,n}^2}^\infty \mathbb{P}\left(n\sum_{j=s+1}^{p}\widehat{\theta}_j^2 > F_\alpha - (1+\sqrt{p})\sum_{j=1}^{s}x_j,\ n\widehat{\theta}_1^2 = x_1, \ldots, n\widehat{\theta}_s^2 = x_s\right) dx_1 \ldots dx_s.$$

Now, notice that the first event in the intersection of events, $n\sum_{j=s+1}^{p}\widehat{\theta}_j^2 > F_\alpha - (1+\sqrt{p})\sum_{j=1}^{s}x_j$, only depends on $n\widehat{\theta}_j^2$ with $j \in \{s+1, \ldots, p\}$. However, the last events in

the intersection of events only depend on $n\widehat{\theta}_j^2$ with $j \in \{1, \dots, s\}$. Using independence, we obtain that:

$$\frac{1}{\mathbb{P}(\overline{C}_s)} \int_{\delta_{p,n}^2}^{\infty} \dots \int_{\delta_{p,n}^2}^{\infty} \mathbb{P}\left( n \sum_{j=s+1}^{p} \widehat{\theta}_j^2 > F_\alpha - (1+\sqrt{p}) \sum_{j=1}^{s} x_j,\ n\widehat{\theta}_1^2 = x_1, \dots, n\widehat{\theta}_s^2 = x_s \right) dx_1 \dots dx_s$$

$$= \frac{1}{\mathbb{P}(\overline{C}_s)} \int_{\delta_{p,n}^2}^{\infty} \dots \int_{\delta_{p,n}^2}^{\infty} \mathbb{P}\left( n \sum_{j=s+1}^{p} \widehat{\theta}_j^2 > F_\alpha - (1+\sqrt{p}) \sum_{j=1}^{s} x_j \right) \prod_{j=1}^{s} f_{n\widehat{\theta}_1^2}(x_j)\ dx_1 \dots dx_s,$$

where we used that the joint probability density of $n\widehat{\theta}_1^2, \dots, n\widehat{\theta}_s^2$ is the product of the marginals by independence. Also notice that these components are identically distributed so their pdfs are the same. Moreover, we can also use this i.i.d. assumption to rewrite $\mathbb{P}(\overline{C}_s)$. We find that $\mathbb{P}(\overline{C}_s) = \mathbb{P}(n\widehat{\theta}_1^2 > \delta_{p,n}^2, \dots, n\widehat{\theta}_s^2 > \delta_{p,n}^2) = \overline{F}_{n\widehat{\theta}_1^2}(\delta_{p,n}^2)^s$. We conclude that the conditional probability for the upper bound can be written as:

$$\frac{1}{\overline{F}_{n\widehat{\theta}_1^2}(\delta_{p,n}^2)^s} \int_{\delta_{p,n}^2}^{\infty} \dots \int_{\delta_{p,n}^2}^{\infty} \overline{F}_{\chi^2_{p-s}}\left( F_\alpha - (1+\sqrt{p}) \sum_{j=0}^{s} x_j \right) \prod_{j=1}^{s} f_{n\widehat{\theta}_1^2}(x_j) dx_1 \dots dx_s.$$

□

*Proof of proposition 2.* We will first rewrite the conditional probability for the sake of notational simplicity:

$$\mathbb{P}\left( n \sum_{j=s+1}^{p} \widehat{\theta}_j^2 > C_\alpha \mid n\widehat{\theta}_{s+1}^2 < \delta_{p,n}^2, \dots, n\widehat{\theta}_p^2 < \delta_{p,n}^2 \right) = \mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)},$$

where $A$ denotes the event that $n\sum_{j=s+1}^{p} \widehat{\theta}_j^2 > C_\alpha$ and B denotes $n\widehat{\theta}_{s+1}^2 < \delta_{p,n}^2, \dots, n\widehat{\theta}_p^2 < \delta_{p,n}^2$. To find the probability of the intersection of $A$ and $B$, we use the law of total probability. Applying this law, $\mathbb{P}(B)$ can be rewritten as:

$$\begin{aligned} \mathbb{P}(B) &= \mathbb{P}(A \cap B) \cup \mathbb{P}(A^c \cap B) \\ &= \mathbb{P}(A \cap B) + \mathbb{P}(A^c), \end{aligned}$$

where we used in the last step that the events are mutually exclusive and that $A^c \subset B$ since $A^c$ is the lower left triangle and $B$ is the box itself. Now, using that $\mathbb{P}(A \cap B) = \mathbb{P}(B) - \mathbb{P}(A^c)$, the conditional probability can be rewritten as:

$$\begin{aligned} \mathbb{P}(A \mid B) &= \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \\ &= 1 - \frac{\mathbb{P}\left( n\sum_{j=s+1}^{p} \widehat{\theta}_j^2 < C_\alpha \right)}{\mathbb{P}\left( n\widehat{\theta}_{s+1}^2 < \delta_{p,n}^2, \dots, n\widehat{\theta}_p^2 < \delta_{p,n}^2 \right)} \\ &\overset{\text{i.i.d.}}{=} 1 - \frac{\mathbb{P}\left( n\sum_{j=s+1}^{p} \widehat{\theta}_j^2 < C_\alpha \right)}{\mathbb{P}\left( n\widehat{\theta}_p^2 < \delta_{p,n}^2 \right)^{p-s}} \\ &= 1 - \frac{F_{\chi^2_{p-s}}(C_\alpha)}{F_{\chi_1^2}\left(\delta_{p,n}^2\right)^{p-s}}. \end{aligned}$$

□