Double Descent in Portfolio Management

Songze Yang

May 26, 2023

Abstract

The large portfolio construction is gaining attention after years after years growth of in the finance sector. The number of assets in a portfolio nowadays can easily exceed the sample size so investors are looking for a method to handle the high-dimensional portfolio construction. A recent development, min-norm regression, can handle the case of high-dimensional data, especially in the presence of latent space. The financial data is known to have noise, a fact that facilitates and motivates the application of min-norm regression in this case. In this monograph, we propose a new formulation to enable and explore the usefulness of the min-norm regression in high-dimensional portfolio construction.

1 Introduction

Denote the excess return for a p asset portfolio at time t as $R_t := (r_{t,1}, r_{t,2}, \ldots, r_{t,p})'$, a $p \times 1$ vector for $t \in 1, .., N$. Denote μ as the target excess return of a portfolio, $\mu := (\mu_1, \mu_2, ...\mu_p)'$ which is also a $p \times 1$ vector. A $p \times p$ covariance matrix of excess return is $\Sigma := E(r_t - \mu)(r_t - \mu)'$ and its sample estimate is $\hat{\Sigma} := \frac{1}{n}\Sigma_{t=1}^{n}(r_t - \bar{r})(r_t - \bar{r})'$, where the $\bar{r}$ is the sample mean of excess asset return. The data matrix of excess return is an $n \times p$ matrix denoted as R with iid copies of excess return for p asset portfolio. Denoted the precision matrix for portfolio contains p assets as Θ and its estimate $\hat{\Theta}$. Lastly, denote the over-parametrization ratio $\frac{p}{n} := \gamma$ and it is limiting condition as $n \to \infty$, $p \to \infty$. Lastly, let us call $n > p$ or $\gamma \in (0, 1)$ as the underparametrized case and $n < p$ or $\gamma > 1$ as the overparametrized case.

Motivation to build a large portfolio has gained attention in recent years as globalization and economic growth enlarge investment choices. Referring to the latest financial data, over 58,000 companies have been listed on various indices in the US stock exchange by the end of Q1 2022, not to mention the assets listed on the global finance indies. Diversification, or specifically global diversification, encourages investors to invest internationally and thus migrate the risk and gain higher returns. The challenges in achieving this goal include that the portfolio constructed loses its accuracy when the number of assets grows sufficiently large. Therefore, investors face a novel challenge in building a large portfolio

with a vast number of assets and meanwhile gaining an optimal return with a controllable risk. Still, the traditional portfolio construction strategy cannot handle the case of such high dimensions as nowadays the number of assets in the portfolio can easily expand to the size that is comparable to our sample size or even exceed it.

Literature review on portfolio construction is well-written and rich. The groundbreaking work on the mean-variance analysis by Markowitz, (1952)[12] laid the foundation for portfolio management. By solving a quadratic optimization problem, the investors can maximize their expected return while controlling the risk limit. Despite the enormous power brought by this framework, the solution to the mean-variance portfolio with plugging-in sample estimate is not reliable as the number of assets increases. More specifically, the optimal solution depends on the population parameters, namely population mean and covariance, which are directly observed, so the error introduced by estimating these parameters accumulates and starts to input more uncertainty into our result.

To reduce the error in the estimation approach, Ledoit and Wolf (2017)[10] put forwards a non-linear shrinkage estimate for the covariance. For the minimum variance portfolio, the solution only depends on the population covariance. Ledoit and Wolf (2017)[11] purpose a mean square risk function and a non-linear shrinkage method on the spectrum of the sample covariance matrix to achieve a l2 asymptotic consistency to the population covariance. Although this method achieves much lower risk as it imposes constraints on the population parameters, its risk still violates the risk constraint while does not achieve Sharpe ratio efficiency (Ao et al., 2018)[1].

Another direction to improve the portfolio performance is to translate the optimization problem and thus bypass the estimation of the population parameters. Take the classical sparse estimation (Brodie et al., 2009)[3] as an example. The optimization problem is first formulated into a constrained least square regression, and an l1 LASSO-type sparsity is added to stabilize the result. In spite of the fact that this method reduces the uncertainty, the l1 sparsity will introduce a feature selection and hence reduce the number of assets in our portfolio and thus cannot handle the high-dimensional case. Also, this method enjoys a level of subjectivity in ways that the response variables in this linear regression are a given risk level purposed by the investors.

To achieve the Sharpe ratio optimal, Ao et al., (2018)[1] prove that the optimization problem in the construction of the mean-variance portfolio can be written as an unconstrained regression problem, specifically a maximum-Sharpe-ratio and sparse regression (MAXSER). The response variables are estimated from the sample Sharpe ratio to reduce the level of subjectivity. Also, this novel regression form does not limit the resulting return to a certain risk level and does not require the weight parameter to be normalized in the optimization. In the same sense, this method does not fit into the high-dimensional portfolio

construction as the estimator for the maximum Sharpe ratio is only consistent when the size of our portfolio is insignificant compared with our sample size. By purposing the LASSO type shrinkage method, the assets number is shrunk. This method achieved a higher Sharpe ratio and lower risk compared with the non-linear shrinkage method.

In Caner et al., (2020)[5], a consistent estimator of the maximum Sharpe ratio is purposed for $\gamma \in (0,1) \cup (1,\infty)$. This consistent estimator is built upon the Nodewise regression precision matrix estimator purposed by Callot et al. (2021)[4] and Change et al. (2018)[6]. The consistent estimator for all γ through the domain enables us to investigate large portfolio construction. However, as mentioned in Ao et al., (2018)[1], consistently estimating the coefficients in high-dimensional cases is in general impossible. Thus, there is a need to find ways to shrink the coefficient estimation error and form a feasible investment strategy.

Although the traditional method seems to be defective in overparameterized regimes and needs dimensional reduction, large data science models, such as the neural network or Xgboost, have seen success in dealing with the ultra-high dimension. These models have observed a consistent drop in the estimation error on both the seen data and the unseen data with the growth in dimension. In other words, not only do these models predict well on the seen cases but also do they generalize well on the unseen data[2], a fact that often contradicts the conventional statistical learning wisdom. An interesting phenomenon for these models is that the prediction risk on the unseen data decreases again after the dimension suppresses the sample size, known as a double descent.

Discovered as a mystery in the Simons Institute program on Foundations of Machine Learning in 2017, a simple min-norm linear regression also appears to have a double descent phenomenon. The simple min-norm regression has been an object of popularity and studied by several authors. Hastie (2022)[8] study the property of the min-norm regression under the asymptotic sense backed by the random matrix theory under linear and non-linear data covariance assumptions. In the case of latent space covariance structure, the min-norm regression comes across as an interpolator in the sense that it interpolates the training data and experiences a monotonic decrease in the prediction risk. Hastie (2019) further develops the framework to decompose the prediction risk into bias and variance and thus provides insight into the theoretical explanation of this phenomenon.

The contribution of this monograph includes that it finds a way to construct a large portfolio in high-dimensional cases even when the number of assets in our portfolio overtakes the sample size. The method we provide is the best for predicting the true function and can be generalized to ultra-high dimension cases. We will showcase the simulation result and state the reason behind the simulation output. This expands the work of Ao et al. (2018)[1] to the over-parameterized cases and provides a novel way to achieve high dimension

portfolio management technique.

The Organization for this paper is as follows. In the next section, we introduce our model formulation and propose our estimation approach, where we explain in detail the methodology and model setup. This includes a review of the traditional results in portfolio management, estimation, and the classical result for random matrix theory. Followed by the theory section, we dive into the min-norm regression and provide the existing theory and known results related. Guided by the theory, we provide the simulation result of our formulation and show that our formulation achieves better results than the existing method. Lastly, we perform real data analysis to test our formulation in a real-world setting.

2 Model Formulation

Diversification is an important concept in portfolio construction. It intends to find a weight parameter w to individual assets that construct a portfolio so that the portfolio maximizes expected return and minimizes the risk. Two of the long-standing problems of this type include finding the global minimum variance portfolio (GMV) and mean-variance portfolio.

The ground-breaking mean-variance portfolio optimization problem first put forward by Markowitz (1952) aims to find a weight parameter w that minimizes the variance or risk at an expected return level ρ. It has a simple solution that relies only on the expected mean and covariance matrix of the asset returns.

$$\hat{w} = argmin_w(w'\Sigma w) \ st. \ w'\mu = \rho \ and \ w'1_p = 1 \tag{1}$$

$$\hat{w} = \frac{\Sigma^{-1}\mu}{1^T\Sigma^{-1}\mu} \tag{2}$$

In many cases, the population mean and covariance are not rather obvious in practice, so the classical approach is to use the sample plug-in estimator, namely the sample mean and covariance matrix. By the classical result from random matrix theory, the sample estimate will not converge to its population counterpart in a high-dimensional overparametrized case where the $n \to \infty$, $p \to \infty$ and $\frac{p}{n} \to \gamma > 1$, a fact that is also known as the curse of dimension.

Interpolators, such as the neural work or Xgboost, built on a sufficiently complex structure, seem to perform and generalize well, even with a perfect fit to the noisy data. It is known that overparameterization is necessary for benign overfitting in regards to these interpolators where the test risk decrease again after a traditional bias-variance trade-off. This fact is an intriguing mystery at first sight as it imposes a mismatch with the traditional statistical learning

theory (Bartlett et al., 2020c). The min norm or ridgeless regression is shown to interpolate the data if we assume that the dimension of the data is large enough.

The mean-variance portfolio optimization falls into a least square problem. For this reason, the risk for the portfolio can decrease again in the high-dimensional case and thus open the door to achieving a better expected return for our portfolio. A regression equivalence of mean-variance portfolio optimization is purposed by Brodie et al. (2008). The goal here is to minimize the squared Euclidean distance between a certain risk (return) level and our portfolio return with constrained weight normalized to 1.

$$w = argmin_w E[|\rho - w' R_t|^2] \ st.\ w'\mu = \rho \ and \ w' 1_p = 1 \tag{3}$$

The normalization constraint may not be needed in the optimization as the weight can be re-scaled to 1 afterward. Also, the given risk level ρ is not obvious in practice and the fund managers sometimes only love to overcome an arbitrary threshold of return. Let the expected portfolio return free from a fixed return level $r*$[1] or ρ[3] and for a given risk constraint σ we can have a novel representation of the mean-variance portfolio (Ao et al., 2018)[1]. This is equal to an unconstrained regression representation of the problem.

$$w = argmin_w (w' \Sigma w) \ st.\ w'\mu \geq r* := \sigma\sqrt{\theta} \tag{4}$$

$$argmin_w E(r_c - w' R_t)^2, \ where \ r_c := \frac{1+\theta}{\theta} r^* = \sigma \frac{1+\theta}{\sqrt{\theta}} \tag{5}$$

The risk constraint r_c is a function of the square of maximum Sharpe ratio θ, a scalar that depends on population parameters μ and Σ. Nevertheless, the weight w is not constrained in the above representation, thus we have an out-of-sample Sharpe ratio. By solving the unconstrained optimization problem, plugging in the weight w_{opt} (Equation (4)), we derive the general formula for the maximum Sharpe ratio given any arbitrary weight w_{opt}. This is also the maximum out-of-sample Sharpe ratio as in equation (A.2) in Ao et al., (2018). Let's denote the Sharpe ratio given an arbitrary weight as $s(w)$ so that we have:

$$w_{opt} = \frac{\sigma}{\sqrt{\theta}} \Sigma^{-1} \mu \tag{6}$$

$$s(w_{opt}) = \frac{w'_{opt}\mu}{\sqrt{w'_{opt}\Sigma w_{opt}}} = \frac{\mu' \Sigma^{-1} \mu}{\sqrt{\mu' \Sigma'^{-1} \Sigma \Sigma^{-1} \mu}} \tag{7}$$

Now we propose our estimation for the out-of-sample Sharpe ratio in the sense of out-of-sample performance (see Theory). This estimator occurs to have the best out-of-sample property but is not consistent. Denote the generalized inverse as $(\cdot)^+$. Notice that after we substitute our estimator for the precision matrix, that is the generalized inverse estimator, the maximum out-of-sample Sharpe ratio is

a function of the generalized precision matrix estimator, population mean, and population variance.

$$s(w_{opt})_+ = \frac{\mu'\Sigma^+\mu}{\sqrt{\mu'\Sigma'^+\Sigma\Sigma^+\mu}},\ \gamma \in (0,\infty) \tag{8}$$

This estimator coincides with the plugging-in sample precision matrix estimator for $\gamma \in (0,1)$. We will compare the goodness of prediction between our generalized inverse estimator and the consistent estimator (see below).

To estimate the r_c , Kan and Zhou (2007) purposed a consistent estimator $\hat{\theta}$ for $\gamma \in (0,1)$:

$$\hat{\theta} := \frac{(n-p-2)\hat{\theta_s} - p}{n},\ where\ \hat{\theta_s} = \hat{\mu}'\hat{\Sigma}^{-1}\hat{\mu} \tag{9}$$

The $\hat{\mu}$ and $\hat{\Sigma}$ are the sample estimate of mean and covariance and their product is the sample maximum Sharpe ratio $\hat{\theta}_s$. Ao et al., (2018) further validate that the estimated r_c from the $\hat{\theta}$ is a consistent estimation for $\gamma \in (0,1)$ with the following assumption: 1. The $\hat{\theta}$ is bounded. 2. the return data matrix follows normality assumption as $R_t \sim N(\mu, \Sigma)$. 3. the sample covariance matrix $\hat{\Sigma}$ is invertible so that the our sample size satisfies $n > p$ condition. The proof is left in the Appendix. Assumption 3 implicitly indicates that the dimension of our data is implicitly assumed to have $\gamma \in (0,1)$. Again, followed by the classical result in random matrix theory, the eigenvalue distribution in $\gamma \in (1,\infty)$ is not uniformly distributed so the Σ^{-1} is out of reach and not defined. Also, our data may not follow the normal distribution if our sample size is small. Ao et al., (2018) introduce a sparsity assumption (a LASSO type shrinkage) to the w estimation in this high-dimensional space, corresponding to $l1$ boundedness of the parameter w. This will perform a variable selection and achieve dimensional reduction.

$$w(r_c) := argmin_w \frac{1}{n}\sum_{t=1}^{n}(r_c - w'R_t)^2, st.||w||_1 \leqslant \lambda \tag{10}$$

Thus, the response variable (risk constraint r_c) and a regression technique for cases where $\gamma \in (1,\infty)$ are unknown. In a recent development, Caner et al., (2020) find a consistent estimator for the maximum Sharpe ratio that stems from the precision matrix estimation denoted as $\hat{\Theta}$ by Callot et al. (2019). The main benefit of this method is that this estimator works on non-iid data. There are assumptions made for the population covariance matrix which needs to satisfy the following assumption: The row in $n \times p$ matrix R needs to have strictly stationary β mixing rows with β mixing coefficients satisfying $\beta_k \leq exp(-K_1 k^B)$ for any positive k, with constants $K_1 > 0, B > 0$ that are independent of n and p.

1. The eigenvalue of the population covariance matrix Σ needs to be bound away from 0 and infinity.

2. There exists constants $K_2 > 0$, $K_3 > 1$ and $B_2, B_3 \in (0, 2]$ that are independent of p and n such that:

$$max_{j\in[1,p]}E(exp(K_2|r_{tj}*)|^{B_2}] < K_3$$
$$max_{j\in[1,p]}E(exp(K_2|\eta_{tj}*)|^{B_2}] < K_3$$
$$\eta_{tj}(residual) := R_{tj} - R_{t,-j}\gamma_j$$

We will rewrite the estimation process from Caner et al., (2020) and the steps are summarized below.

1. Estimate $\hat{\gamma_j}$ for a given λ_j by solving:

$$\hat{\gamma_j} := argmin_{\gamma\in R^{p-1}}(||R_{t,j} - R_{t,-j}\gamma||_2^2/n + 2\lambda_j||\gamma||_1)$$

The γ_j here is a $p-1\times 1$ regression coefficient and the $R_{t,-j}$ denotes the excess return at time t from portfolio j removes the j^{th} element from row $R_{t,j}$. Noted that our row and column representation are not conformed to normal representation.

2. Find the tuning parameter λ_j by the GIC information criterion of Fan and Tang. (2013)[7]:

$$GIC(\lambda_j) := log(\hat{\sigma^2_{\lambda_j}}) + |\hat{S}_j(\lambda_j)|\frac{log(p)}{n}log(log(n))$$

3. Repeat step 1 and 2 for $j = 1, \ldots, p$.

4. Compute the $\hat{C}$ and $\hat{T}^2$, where $\hat{C}$ is a square matrix with identity diagonal defined as:

$$\begin{bmatrix} 1 - \hat{\gamma}_{1,2}... - \hat{\gamma}_{1,p} \\ -\hat{\gamma}_{2,1}1... - \hat{\gamma}_{2,p} \\ .....1... \\ -\hat{\gamma}_{p,1} - \hat{\gamma}_{p,2}...1 \end{bmatrix}$$

T^2 are defined as:

$$T^2 := diag(\tau_1^2, ..., \tau_p^2)$$
$$\tau_j^2 = \frac{1}{\Theta_{jj}}$$

Its estimation $\hat{T}^2$ are defined as:

$$\hat{T}^2 := diag(\hat{\tau}_1^2, ..., \hat{\tau}_p^2)\hat{\tau}_j := \frac{||R_tj - R_{t,-j}\hat{\gamma}_j||_2^2}{n} + \lambda_j||\hat{\gamma}_j||_1$$

5. Return the precision matrix by $\hat{\Theta} = \hat{T}^{-2}\hat{C}$.

Ao et al., (2018) prove that the maximum out-of-sample Sharpe ratio with the sample mean and sample variance estimator is not consistent in the under-parametrized case so the estimator in Kan and Zhou (2008) is considered. Caner

et al., (2020) continue their work and purpose an estimator for weight w_{opt} that plugs in the Nodewise regression estimator for the precision matrix. Here, the Θ is the precision matrix and its Nodewise estimator is $\hat{\Theta}$. Therefore, we achieve the maximum out-of-sample Sharpe ratio with this estimator for weight w_{opt}:

$$\hat{w}_{opt} = \frac{\sigma\hat{\Theta}\hat{\mu}}{\sqrt{\hat{\mu}\hat{\Theta}\hat{\mu}}} \tag{11}$$

$$\hat{s}(\hat{w}_{opt}) = \frac{\hat{w}'_{opt}\mu}{\sqrt{\hat{w}'_{opt}\Sigma\hat{w}_{opt}}} = \frac{\hat{\mu}'\hat{\Theta}\mu}{\sqrt{\hat{\mu}'\hat{\Theta}'\Sigma\hat{\Theta}\hat{\mu}}} \tag{12}$$

Notice that by the formula from weight $w_{opt}\Sigma w_{opt} = \sigma^2$ and this leaves us to consider how good are the $\hat{w}'_{opt}\mu$ with respect to the population weighted mean $w'_{opt}\mu$ and the $\hat{w}_{opt}\Sigma\hat{w}_{opt}$ with respect to the population variance σ^2 . It proves that these two conditions will both converge with a rate of $o_p(1)$ for $\gamma > 1$ (Caner et al., 2020).

$$|\frac{w'_{opt}\mu}{\hat{w}'_{opt}\mu} - 1| = O_p(\bar{s}\sqrt{lnp/n}) = o_p(1)$$

$$|\hat{w}_{opt}\Sigma\hat{w}_{opt} - \sigma^2| = O_p(\bar{s}\sqrt{lnp/n}) = o_p(1)$$

Also, the maximum out-of-sample Sharpe ratio ($MOOSSR$) will also converge to the maximum Sharpe ratio (MSR) which is proofed in Caner et al., (2020) at the same rate above for $\gamma > 1$.

$$|(\frac{MOOSSR}{MSR})^2 - 1| = O_p(\bar{s}\sqrt{lnp/n}) = o_p(1)$$

Therefore, we have our maximum out-of-sample Sharpe ratio ($MOOSSR$) as a estimator, in case of an unconstrained weight, for maximum Sharpe ratio (MSR). This estimator is only good for $\gamma > 1$ and this shall use as a complimentary for case where $\gamma \in (0, 1)$. The overall formula for the square of maximum Sharpe ratio is illustrated below. To this point, we have acquired all the elements we need to calculate the r_c in all cases.

$$\hat{\theta} = \begin{cases} (\frac{\hat{\mu}'\Theta^{-1}\mu}{\sqrt{\hat{\mu}'\Theta'\Sigma\Theta\hat{\mu}}})^2, & \gamma \geq 1 \\ \frac{(n-p-2)\hat{\theta}_s - p}{n}, & \gamma \in (0,1) \end{cases}$$

$$r_c := \sigma\frac{1+\hat{\theta}}{\sqrt{\hat{\theta}}} \tag{13}$$

The work left is to formulate a high-dimension regression to enable us to work in overparametrized case. The least square problem for mean variance optimization by Ao et al. (2018) introduces a linear form. We will restate our notation here. The risk constrain r_c is a scalar and it is a function of maximum Sharpe ratio so the true r_c depends only on μ and Σ and does not depend on time t. The R_t

is a $1 \times p$ row vector from the data matrix R and the weight is a $p \times 1$ column vector. Let's bring in the error term denoted as ϵ_i for i = 1,...,n observations. We will assume the error term and the data follow some unspecific distribution denoted as P_{R_t} which has mean μ and covariance Σ and P_{ϵ_i} on R^p dimensional space with mean 0 and variance σ^2.

$$r_c = w'R_t + \epsilon_i, (R_t, \epsilon_i) \sim P_{R_t} \times P_{\epsilon_i} \tag{14}$$

Empirically, the training risk function for the min norm regression will be the sample average of square distance between the true return and the predictive return. The sparsity term is reasonably excluded as we do not need to do the dimension reduction. By solving the min-norm regression, we can get the optimal weight solution in all $\gamma \in (0, \infty)$ cases. Denote the generalized inverse as $(\cdot)^+$. To make a prediction on new return R_{test}, we can utilize the optimal parameter w. Denote the predictive portfolio return on the new portfolio as r_{pred} then:

$$risk_{train} := \frac{1}{n}\sum_{t=1}^{n}(r_c - w'R_t)^2,\ t = 1, ..., n \tag{15}$$

$$w(r_c) := argmin_w \frac{1}{n}\sum_{t=1}^{n}(r_c - w'R_t)^2 = r_c R_t'(R_t R_t')^+ \tag{16}$$

$$r_{pred} := w'(r_c)R_{t_{test}} \tag{17}$$

To evaluate the performance of our fit, let's put forward the empirical mean square error as test criteria. The empirical test mean the square error will approximate the true MSE when the sample size n is sufficiently large.

$$risk_{test} := \frac{1}{n}\sum_{t=1}^{n}(r_c - r_{pred})^2 \approx E(r_c - w'R_t)^2,\ t = 1, ..., n \tag{18}$$

Beyond the traditional bias-variance trade-off in the underparameterized case, the min norm regression enables us to work in the overparametrized case, which is used to consider as singular. Built upon the traditional bias-variance trade-off, the local minimum in the underparametrized case is widely known but a global minimum may exist further in the overparametrized case. Many interpolators show a double or single descent in the overparametrized case which promotes the importance to investigate the overparametrized case.

3 Theory

The min norm estimator has a simple definition that the min norm optimization finds the $p \times 1$ optimal parameter $\hat{w}$ that solves a constrained optimization problem below given $n \times p$ data matrix R and response r_c as $n \times 1$ return

vector. Denote the $||\cdot||$ norm here to be the Euclidean in R^n or the Hilbert space norm as the min norm regression can generalize to the infinite-dimensional case. We can rewrite the risk function in matrix form and get the optimization problem:

$$\hat{w} = argmin||w||_2 : w \ minimizes \ ||r_c - Rw||_2^2 \tag{19}$$

In practice, the expected loss function is approximated by its averaged sample version when the sample size is sufficiently large.

$$risk := \frac{1}{n}(r_c - Rw)^2 \tag{20}$$

From another point, it can also be viewed as ridgeless regression in comparison to ridge regression estimator $\hat{w}_\lambda$ when the penalty term λ approaches zero. The $\hat{w}$ approaches $\hat{w}_\lambda$ as the $\lambda \to 0$. This can be seen clearly from the analytical solution for the ridge regression. We use I to denote the identity operator on H or on a finite $n \times n$ identity matrix.

$$\hat{w}_\lambda = argmin_{w \in R^p} \frac{1}{n}||r_c - Rw||_2^2 + \lambda||w||_2^2 \tag{21}$$

In the limiting condition, the true inverse will equal to the generalized inverse of the linear operator $R'R$, which guarantees that the $R'R$ is bounded and has a closed range (Bartlett et al., 2020b).

$$\hat{w} \to \hat{w}_\lambda, \ as \ (R'R + n\lambda I)^{-1}R'r_c \to (R'R)^+R'r_c \tag{22}$$

In the underparameterized case, the solution to the min-norm regression coincides with simple linear regression, whose solution is unique. In the overparametrized case, simple linear regression has many solutions but the min norm solution gives the minimized $l2$ distance for $||r_c - Rw||^2$ in the separable Hilbert space $H[2]$, a fact that enables more flexibility and extends the simple linear regression to ultra-high dimension. The case in the finite space R^n is only a generalization.

Not as the underparameterized case where the test risk only depends on the variance of the model, it is shown that the test risk of min-norm regression depends on the geometry of the covariance on the data matrix and the w (Hastie, 2019), denoted as pair (Σ, w). Without loss of generality assume that $E(R_t) = 0$ and $Cov(R_t) = \Sigma$ as one can always demean and scale variance to get any kind of distribution. Thus, safely assume that our data point R_t has covariance Σ and 0 mean. The eigenspectrum of Σ can be written in $\sum_{i=1}^{p} s_i v_i v_i'$, where s_i is the eigenvalue and the v_i is the eigenvector. Therefore, the geometry is impacted by all the s_i and the projection of w to each of the eigenvectors $(v_1w_1, v_2w_2, .., v_pw_p)$. Assume $R_t = \Sigma^{1/2}z_i$ a positive definite Σ covariance and depends on some iid factors z_i with mean 0 and unit covariance without loss of generality. It can be shown that this general form still applies when the R_t and

z_i have a non-linear relationship[8].

There are two components, namely the bias and variance in regards to model parameter w, that play an important role in the decomposition of the risk of the min norm estimator. In the presence of the least square risk, the bias and variance are guaranteed.

$$B_R(\hat{w}; w) = w'\Pi\Sigma\Pi w \; and \; V_R(\hat{w}; w) = \frac{\sigma^2}{n} Tr(\hat{\Sigma}^+\Sigma) \tag{23}$$

where the $\hat{\Sigma} = R'R/n$ is the uncentered sample covariance of R, and $\Pi = I - \hat{\Sigma}^+\hat{\Sigma}$ is the projection to the null space of R.

Now we can show the result in the underparametrized case. Follow the known results in random matrix theory and assume a finite 4th moment for Σ. In the limiting condition as $n \to \infty$, $p \to \infty$ and $\gamma \in (0, 1)$, the risk of min norm estimator satisfies, almost surely:

$$lim_{n\to\infty} risk_R(\hat{w}; w) = \sigma^2 \frac{\gamma}{1-\gamma} \tag{24}$$

The min norm regression experiences the classical bias-variance trade-off in the underparametrized case. The risk, in this case, is just the variance as the Π in bias is 0 given that $\hat{\Sigma}^+\hat{\Sigma} = \hat{\Sigma}^{-1}\hat{\Sigma} = I$.

Nevertheless, This is not the case in the overparametrized case. Safely assume our data follow $R_t = Iz_i$ or any structure that is sphere-like, that is, in summary, the Isotropic case. It can be shown that the risk has a global minimum in the underparametrized case. For Anisotropic cases where the eigenvalue of the covariance matrix is full rank but not equally distributed in every direction, the result is similar. Intuitively, if an increase in feature brings an additional unknown dimension for any fixed n, then the new dimension does not provide much information.

However, if the weight is only associated with a subset of dimension (the subset of basis) of covariance Σ or the increase of feature brings new information and no new dimension, it is shown that the risk is monotone decreasing, a fact matches the performance of the large model as neural network or Xgboost. A similar result is also shown by Bartlett et al. (2020b)[2].

Many works of literature have documented that the finance return data is often accompanied by noise. It is widely studied that the subspace models, such as the factor model, which does dimension reduction and returns the low dimensional latent factors, work well for financial data. If the assumption meets the practice, the risk will decrease monotonically while the return on the unseen portfolio return will be maximized given the risk constraint. We will simulate these cases in the following part.

4 Simulation

Denote μ as the population mean, Σ as the population covariance matrix, and $\hat{\Sigma}$ as the sample covariance matrix. The precision matrix and its Nodewise estimation are denoted as Θ and $\hat{\Theta}$. The square of the maximum Sharpe ratio and its estimation are denoted as θ and $\hat{\theta}$.

For simplicity, we will assume the $n \times p$ data matrix. The weight vector w with dimension $p \times 1$ and the target variable r_c with dimension $n \times 1$. The process of our data-generating process is defined as follows:

1. Generate a positive definite population covariance matrix Σ and:

1. For the isotropic case, covariance matrix Σ will be the identity matrix I and its scale.

2. For the anisotropic case, covariance matrix Σ will equal some positive definite matrix generated by QR decomposition. In the decomposition, it will produce an orthonormal matrix Q and an upper triangular matrix R. Then, we will generate a diagonal matrix D as an approximate to the middle diagonal matrix in the SVD, in this way, we can control the eigenvalue of Σ to bound between finite values, eg. in (0,1). Then, we will equal $\Sigma = QDQ'$.

3. For latent space case, covariance matrix Σ is equal to some lower rank matrix and will also be generated by QR decomposition. But when we generate the middle diagonal matrix, we will assume a part of the eigenvalue is 0 so that the true data covariance only depends on the top a few eigenvalues. Then, we will equal $\Sigma = QDQ'$.

2. Choose a test data sample size l and generate test and train data by multiplying iid norm distributed vector z_i with the population covariance matrix. To make our covariates follow a correlation structure correspond to Σ, we will do as follow: $x_i = z_i * \Sigma^{1/2}$, assuming the z_i is a $1 \times p$ vector and the Σ is the $p \times p$ matrix. By definition, our μ will equal the $z_{mean} * \Sigma^{1/2}$. Assume $z_i \sim N(\mu, \sigma = 1)$ with some arbitrary mean μ and the unit variance.

3. Calculate the maximum Sharpe ratio estimator. We will compare our generalized inverse estimator and the consistent estimator proposed by Kan and Zhou (2007) [9] in underparameterized case or the Nodewise regression estimator proposed by Caner et al., (2020)[4]in overparameterized case following the steps as in model formulation.

4. Calculate the r_c as the target variable in our min-norm regression. As the optimal portfolio will not change with respect to individual asset, the target variable is always the same at each time t.

5. Fit a min-norm regression and return the test MSE risk on a sufficiently large test size l.

The above data-generating process is going to enjoy much flexibility and enable us to explore the relationship with different geometric structures in the data covariance matrix and weight parameter w. We will consider three cases: the isotropic, anisotropic, and latent space cases. The expectations for isotropic and anisotropic cases are that the local minimum will be placed in the underparameterized case but the prediction accuracy will decline again after the overparameterized limit, where $\gamma = 1$, and then ascend. In the above three cases, the risk in the underparameterized case is just the variance of the error terms in our model, which is not observed, while the risk is both bias and variance in the overparameterized case, which is much more complicated depending on Σ and w. In the latent space, prediction accuracy can be more difficult to anticipate as it depends on the relative magnitude of the variance in underparameterized case compared with the sum of variance and bias in the overparameterized case, however, the theory suggests a monotonic decrease in the overparameterized case.

The simulation result for the isotropic case shows a double descent for all estimators (see Figure 1). A large ascent of test MSE at overparameterized limit $\gamma = 1$. A close second look at the graph illustrates that although the test MSE declines largely after the overparameterized limit, it sooner and later soars much larger level compared to the underparameterized case for the consistent estimator (see Figure 2). For these consistent estimators, the global minimum is achieved in underparameterized cases. However, the global minimum for the generalized inverse estimators is achieved in the overparameterized case. Despite some random spike, the out-of-sample performance in the overparameterized overtakes that of the Nodewise estimator. For any fixed n in this case, the feature is not correlated and equally distributed on each of the dimensions, hence the increase in the dimension is not in favor of the prediction. In the underparameterized case, our generalized inverse estimator coincides with the sample out-of-sample estimator, but the out-of-sample performance for the generalized inverse estimator is comparable to the Kan and Zhou (2007) estimator. In addition, the global minimum achieved by combining the generalized inverse estimator and the Nodewise estimator is slightly lower than combination of Kan and Zhou (2007) and the Nodewise estimator (see Figure 3).

Considering the anisotropic case, it also shows a double descent for all estimators (see Figure 4). For consistent estimators, a Zoom-in picture at the graph illustrates that although the test MSE declines largely after the overparameterized limit, it sooner and later soars much larger level compared to the underparameterized case for the consistent estimator (see Figure 5). For these consistent estimators, the global minimum is achieved in underparameterized cases. However, the global minimum for the generalized inverse estimators is achieved in the overparameterized case. Despite some random spike, the out-

of-sample performance in the overparameterized overtakes that of the Nodewise estimator even in the anisotropic case. In the anisotropic underparameterized case, our generalized inverse estimator coincides with the sample out-of-sample estimator, but the out-of-sample performance for the generalized inverse estimator suppresses the Kan and Zhou (2007) estimator (see Figure 6). In addition, the global minimum achieved by combining the generalized inverse estimator and the Nodewise estimator is slightly lower than the combination of Kan and Zhou (2007) and the Nodewise estimator. In the anisotropic case, the generalized inverse estimator outperforms both Kan and Zhou (2007) and the Nodewise estimator in both under and over-parameterized cases.

The target variable in our case is the risk constraint r_c, which is a function of the maximum Sharpe ratio and the given risk level σ. In theory, the magnitude of the σ will affect the target and thus enlarge the estimated weight as $w' = (R'R)^+R'r_c$. Meanwhile, the population mean return μ will affect the return matrix, and thus grow the test MSE in our min-norm regression. In the anisotropic case, we will look at the effect of the population mean μ and the given risk level σ (see graph below). A double descent exists as expected for different μ and σ levels. The huge gap between the $\gamma = 1$ and the others reflect the singularity at this point. The overview figure is similar to that of the isotropic cases. Excluding the singular point, we can spot that both the μ and σ are positively related to the test MSE. This confirms the theory that the test risk of min-norm regression depends on the geometry of weight as both μ and σ that enlarge the w.

Lastly, we present the most important latent space cases. The finance data is extremely noisy in ways that it can be decomposed into a subspace. The design of Σ here is that our return data will only be correlated with a number of top eigenvalues, and thus the true geometry of our return data will only be correlated with a number of top eigenvalues that are fixed.

Concerning the latent space case, it also shows a double descent for all estimators (see Figure 7). For consistent estimators, a Zoom-in picture at the graph illustrates that although the test MSE declines largely after the overparameterized limit, it sooner and later soars much larger level compared to the underparameterized case (see Figure 8). For these consistent estimators, the global minimum is achieved in underparameterized cases. However, the global minimum for the generalized inverse estimators is achieved in the overparameterized case. Despite some random spike, the out-of-sample performance in the overparameterized overtakes that of the Nodewise estimator even in this case. In the underparameterized case, our generalized inverse estimator coincides with the sample out-of-sample estimator, but the out-of-sample performance for the generalized inverse estimator is at a similar level to the Kan and Zhou (2007) estimator (see Figure 9). In addition, the global minimum achieved by combining the generalized inverse estimator and the Nodewise estimator is slightly lower than the combination of Kan and Zhou (2007) and the Nodewise esti-

mator. In the latent space case, the generalized inverse estimator outperforms both Kan and Zhou (2007) and the Nodewise estimator in both under and over-parameterized cases.

The main reason behind our latent case simulation is that the portfolio construction is inherently an unsupervised learning problem where the target variables, the risk constraint, is not observed. Hence, the variance of the model or error ($\epsilon \sim (0, \sigma^2)$ is not directly observed. In cases where the error volatility is larger than the risk in the overparameterized case, the global minimum is expected to be placed in the overparameterized case. Otherwise, the global minimum is placed in the overparameterized case. The consistent estimator is good in the asymptotic sense but the estimation error accumulates, thus the performance gets worse as the growth of dimension.

From all the Figures, it is obvious that our min norm estimator experiences double descent. The test mean square error increases sharply around the $\gamma = 1$ point and then decreases again after that. For $\gamma \in (0, 1)$, we have a bias and variance trade-off as in classical statistics theory. For $\gamma > 1$, we have the double descent on test data, and the test risk after $\gamma = 1$ will depend on the geometry of the w and the Σ. Our proposed generalized inverse estimator is consistently better than the other estimators. The global minimum in overparameterized cases is only achieved for our generalized inverse estimators.

The result is accorded to the theory (Hastie, 2019)[8]. The result shows that as long as the result is not too far away from the interpolation boundary, the test risk decreases to a level similar to the underparametrized case.

5 Conclusion

In this monograph, I propose a method that extends the portfolio construction problem for the mean-variance portfolio to the ultra-high dimension. The method appears excellent in the simulation result. By utilizing the min-norm regression in the mean-variance portfolio with our proposed maximum Sharpe ratio estimator, investors can achieve a much better return in the high dimension even on the unseen data.

References

[1] Mengmeng Ao, Li Yingying, and Xinghua Zheng. Approaching mean-variance efficiency for large portfolios. *The Review of Financial Studies*, 32(7):2890–2919, 2019.

[2] L Bartlett Peter and M Long Philip. Lugosi gábor, and tsigler alexander,"benign overfitting in linear regression,". *Proceedings of the National Academy of Sciences*, 117:30063–30070, 2020.

[3] Joshua Brodie, Ingrid Daubechies, Christine De Mol, Domenico Giannone, and Ignace Loris. Sparse and stable markowitz portfolios. *Proceedings of the National Academy of Sciences*, 106(30):12267–12272, 2009.

[4] Laurent Callot, Mehmet Caner, A Özlem Önder, and Esra Ulaşan. A nodewise regression approach to estimating large portfolios. *Journal of Business & Economic Statistics*, 39(2):520–531, 2021.

[5] Mehmet Caner, Marcelo Medeiros, and Gabriel FR Vasconcelos. Sharpe ratio in high dimensions: Cases of maximum out of sample, constrained maximum, and optimal portfolio choice. 2020.

[6] Jinyuan Chang, Yumou Qiu, Qiwei Yao, and Tao Zou. Confidence regions for entries of a large precision matrix. *Journal of Econometrics*, 206(1):57–82, 2018.

[7] Yingying Fan and Cheng Yong Tang. Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society: SERIES B: Statistical Methodology*, pages 531–552, 2013.

[8] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949–986, 2022.

[9] Raymond Kan and Guofu Zhou. Optimal portfolio choice with parameter uncertainty. *Journal of Financial and Quantitative Analysis*, 42(3):621–656, 2007.

[10] Olivier Ledoit and Michael Wolf. Nonlinear shrinkage of the covariance matrix for portfolio selection: Markowitz meets goldilocks. *The Review of Financial Studies*, 30(12):4349–4388, 2017.

[11] Olivier Ledoit and Michael Wolf. The power of (non-) linear shrinking: A review and guide to covariance matrix estimation. *Journal of Financial Econometrics*, 20(1):187–218, 2022.

[12] Harry M Markowitz. Foundations of portfolio theory. *The journal of finance*, 46(2):469–477, 1991.

6 Appendix

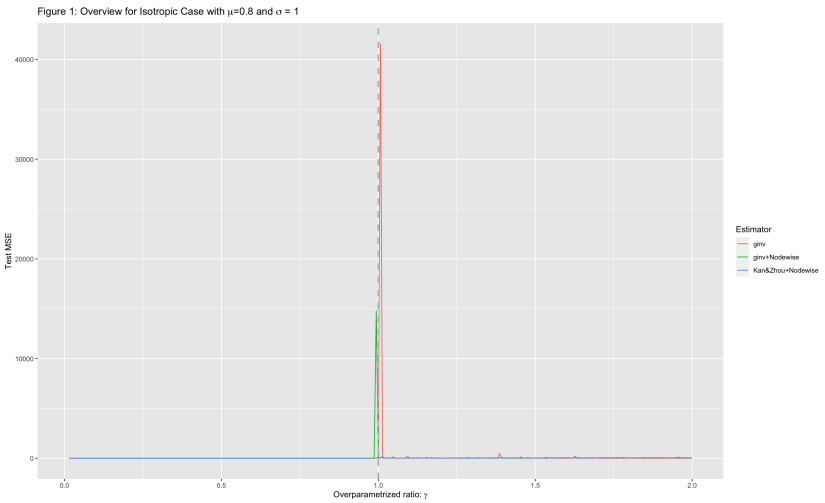

Figure 1:

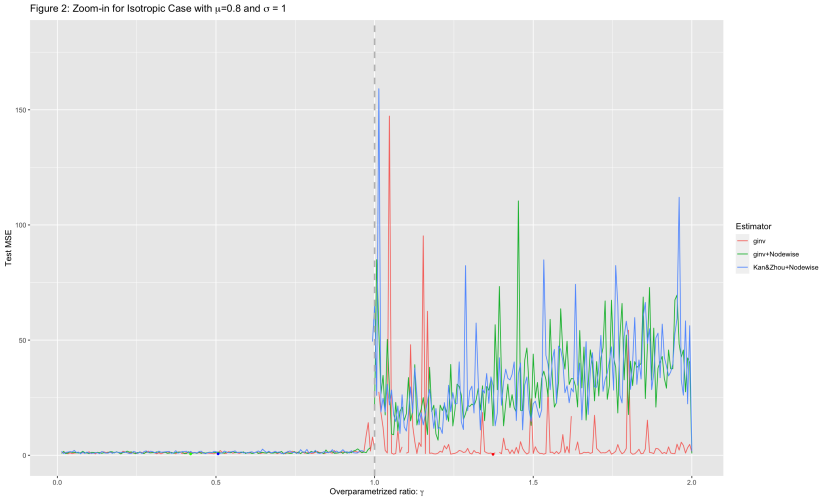

Figure 2:

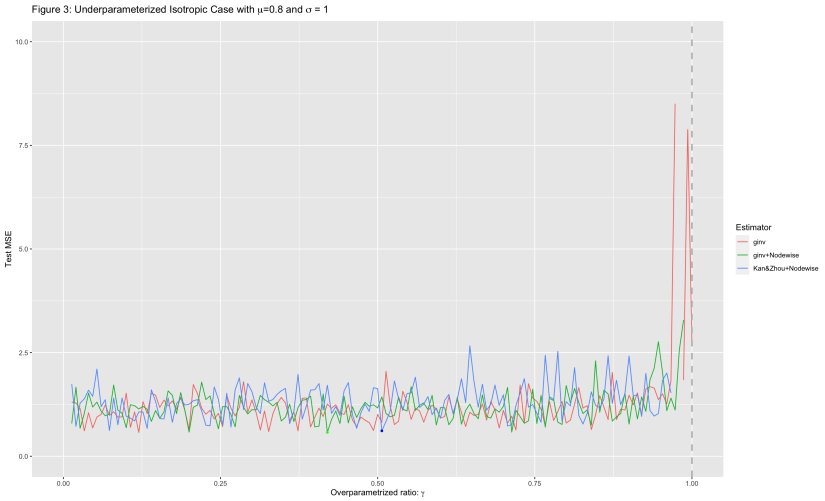

Figure 3:

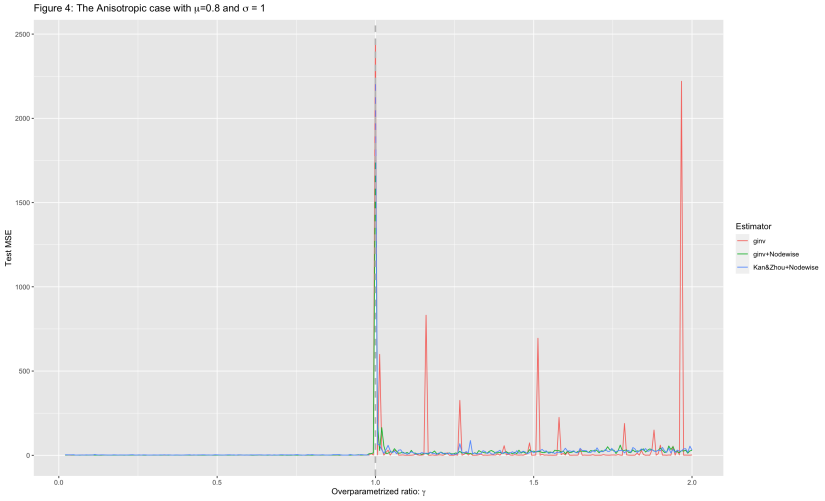

Figure 4:

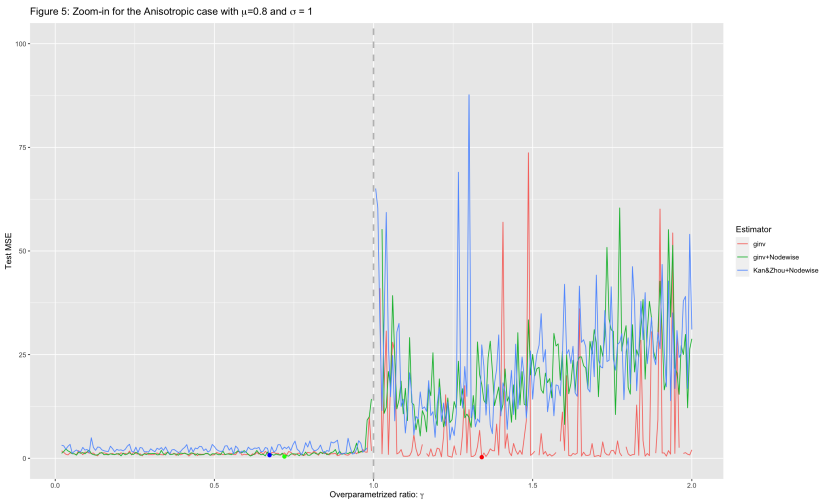

Figure 5:

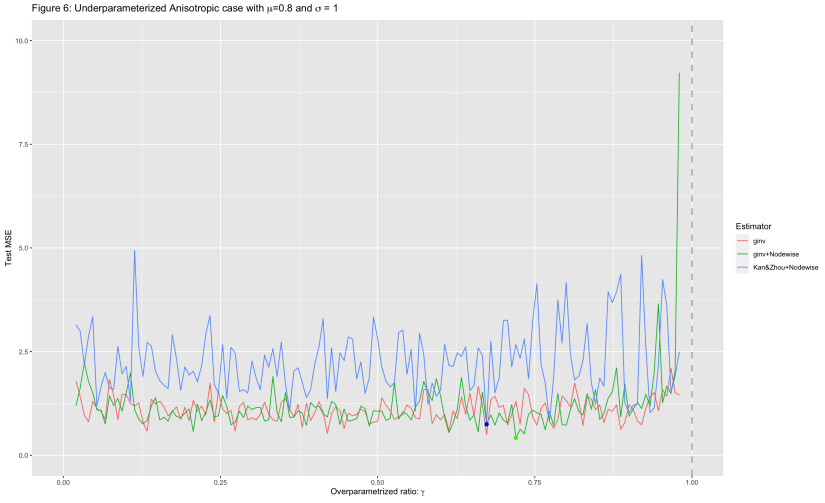

Figure 6:

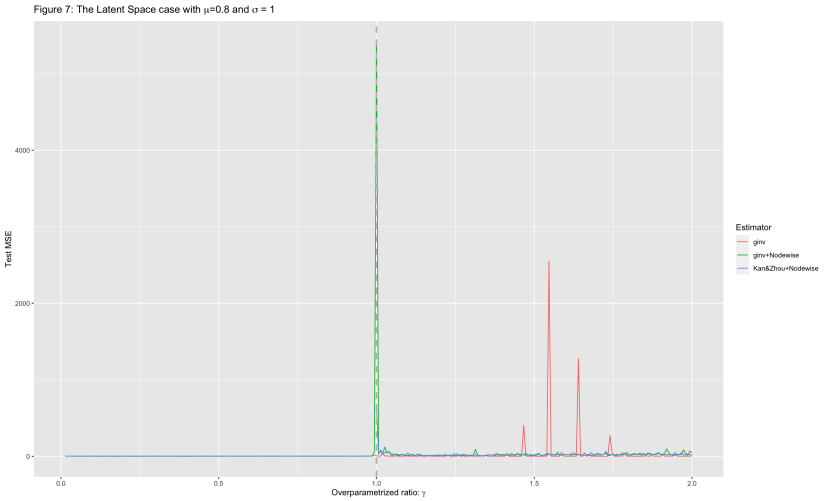

Figure 7:

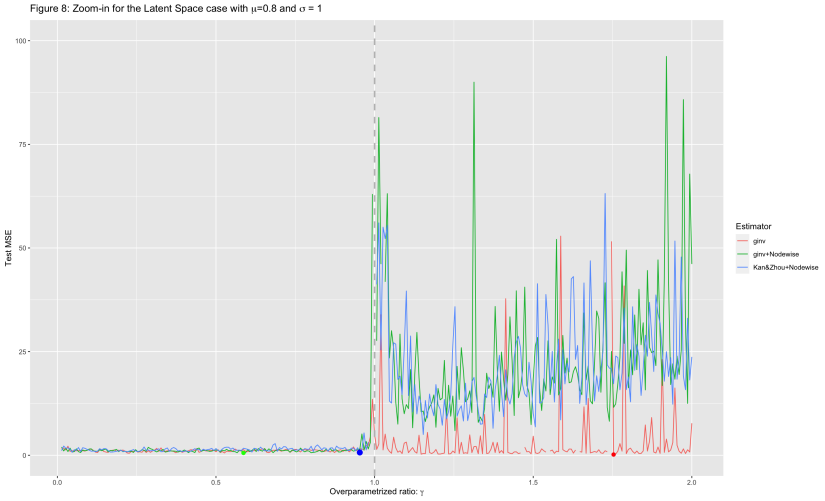

Figure 8:

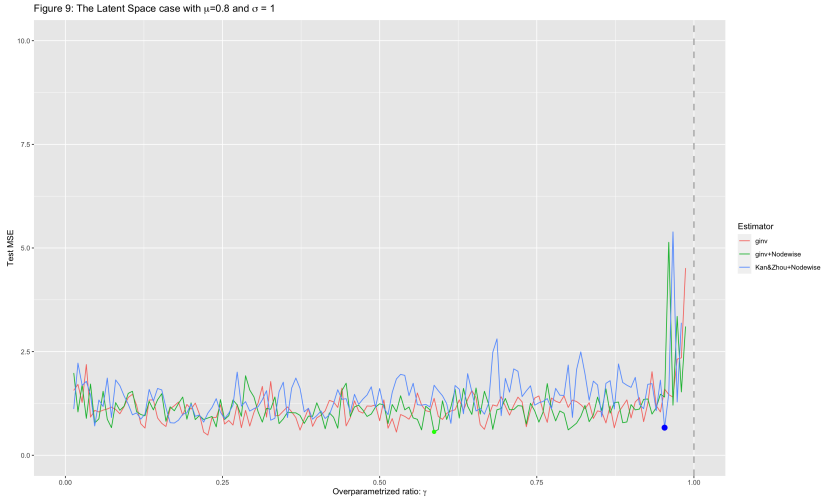
Figure 9: The Latent Space case with μ=0.8 and σ = 1
Test MSE
10.0
7.5
5.0
2.5
0.0
0.00
0.25
0.50
0.75
1.00
Overparametrized ratio: γ
Estimator
ginv
ginv+Nodewise
Kan&Zhou+Nodewise

Figure 9: