BIO 235/ BST 235 Fall 2019

Project: Functional Estimation in High Dimensional Problems

Due: December 18 *Arin Madenci, Tudor Cristea-Platon*

1 Project statement

The purpose of this project is to study estimation of functionals such as the noise variance and signal strength in high dimensional problems.

We will consider the Gaussian Linear Regression Model with random design. Mathematically this means that the data is given by $(y_i, \mathbf{x}_i)$, with $y_i \in \mathbb{R}$ and $\mathbf{x}_i \in \mathbb{R}^d$ where $i \in 1, 2, \ldots, n$. We will assume that

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i, \quad \mathbf{x}_i \sim N_d(\mathbf{0}, \Sigma), \quad \epsilon_i \sim N(0, \sigma^2)$$

which can also be written more concisely as $\boldsymbol{y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$.

We define the following quantities of interest:

- Noise Variance: σ^2
- Signal Strength: $||\Sigma^{1/2}\boldsymbol{\beta}||^2 = \boldsymbol{\beta}^T \Sigma \boldsymbol{\beta}$

One of the main questions we will try to answer in throughout this report is how do we estimate the previously mentioned quantities of interest when the $d \to \infty$ and $n \to \infty$.

2 Regime 1, $\frac{d}{n} \to \rho = 0$

Consider the case where $\frac{d}{n} \to 0$. Without loss of generality, we make the simplifying assumption that $\Sigma = I$. Note in the general case, we only need to replace $(\mathbb{X}, \boldsymbol{\beta})$ by $(\mathbb{X}\Sigma^{-1/2}, \Sigma^{1/2}\boldsymbol{\beta})$. Let us first focus on the estimation of σ^2. We saw in class that $\widehat{\boldsymbol{\beta}}_{MLE} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\boldsymbol{y}$ is MLE estimator of $\boldsymbol{\beta}$ and $\widehat{\sigma}^2_{MLE} = \frac{1}{n}||P_{\mathcal{C}(\mathbb{X})^\perp}(\mathbf{y})||^2$ is the MLE estimator of σ^2. More specifically $\widehat{\sigma}^2_{MLE} = \frac{\boldsymbol{\epsilon}^T P_{\mathcal{C}(\mathbb{X})^\perp} \boldsymbol{\epsilon}}{n} = \frac{\sigma^2}{n} \left(\frac{\boldsymbol{\epsilon}}{\sigma}\right)^T P_{\mathcal{C}(\mathbb{X})^\perp} \left(\frac{\boldsymbol{\epsilon}}{\sigma}\right)$.

We note that $\frac{\sigma^2}{n} \left(\frac{\boldsymbol{\epsilon}}{\sigma}\right)^T P_{\mathcal{C}(\mathbb{X})^\perp} \left(\frac{\boldsymbol{\epsilon}}{\sigma}\right) \Big| \mathbb{X} \sim \chi^2_{n-\mathrm{rank}(\mathbb{X})}(0)$. From [1], we know that for $d \leq n$, $P(\mathrm{rank}(\mathbb{X} = d)) = 1$ as $\mathbf{x}_i$ is continuous. This implies that

$$\boxed{\frac{\sigma^2}{n} \left(\frac{\boldsymbol{\epsilon}}{\sigma}\right)^T P_{\mathcal{C}(\mathbb{X})^\perp} \left(\frac{\boldsymbol{\epsilon}}{\sigma}\right) \Bigg| \mathbb{X} \sim \chi^2_{n-d}(0)}$$

$$\mathbb{E}[\widehat{\sigma}^2_{MLE}] = \frac{\sigma^2}{n}(n-d) \to \sigma^2, \quad \text{as } n \to \infty \text{ s.t. } n - d \to \infty$$

$$\mathbb{V}ar[\hat{\sigma}^2_{MLE}] = \left(\frac{\sigma^2}{n}\right)^2 2(n-d) \to 0, \quad \text{as } n \to \infty$$

The signal strength is given by

$$\begin{aligned}
\mathbb{E}[\hat{\tau}^2_{MLE}] &= \mathbb{E}||\hat{\boldsymbol{\beta}}^2_{MLE}||_2^2 \\
&= \mathbb{E}||(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\boldsymbol{y}||_2^2 \\
&= \mathbb{E}||(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T(\mathbb{X}\boldsymbol{\beta}+\boldsymbol{\epsilon})||_2^2 \\
&= \mathbb{E}||\boldsymbol{\beta} + (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\boldsymbol{\epsilon}||_2^2 \\
&= \mathbb{E}||\boldsymbol{\beta}||_2^2 + \mathbb{E}||(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\boldsymbol{\epsilon}||_2^2 \\
&= \mathbb{E}[\tau^2] + \mathbb{E}||(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\boldsymbol{\epsilon}||_2^2
\end{aligned}$$

The bias of the signal strength can be calculated as

$$\begin{aligned}
\mathbb{E}||(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\boldsymbol{\epsilon}||_2^2 &= \mathbb{E}[\boldsymbol{\epsilon}^T\mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\boldsymbol{\epsilon}] \\
&= \mathbb{E}[\boldsymbol{\epsilon}^T\mathbb{X}(\mathbb{X}^T\mathbb{X})^{-2}\mathbb{X}^T\boldsymbol{\epsilon}] \\
&= \mathbb{E}[tr(\mathbb{X}(\mathbb{X}^T\mathbb{X})^{-2}\mathbb{X}^T\sigma^2\boldsymbol{I})] \\
&= \sigma^2\mathbb{E}[tr(\mathbb{X}(\mathbb{X}^T\mathbb{X})^{-2}\mathbb{X}^T)] \\
&= \sigma^2\mathbb{E}[tr((\mathbb{X}^T\mathbb{X})^{-1})] \\
&= \frac{\sigma^2}{n}\mathbb{E}[tr\Big(\frac{\mathbb{X}^T\mathbb{X}}{n}\Big)^{-1}] \\
&= \frac{\sigma^2}{n}tr\Big(\frac{\Sigma^{-1}}{n-d-1}\Big)
\end{aligned}$$

As $\Sigma = \boldsymbol{I}$, we obtain the bias of the signal strength to be

$$\mathbb{E}||(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\boldsymbol{\epsilon}||_2^2 = \sigma^2\frac{d}{n(n-d-1)}$$

which in the limit of $d/n \to 0$ goes to zero. We point our that in this regime the MLE estimators for noise variance, σ^2, and signal strength, $\boldsymbol{\beta}$, coincide with OLS estimators. We proceed to investigate these estimators using numerical simulation.

2.1 Numerical simulation

We proceed to investigate these estimators using numerical simulations. Consider, the Monte Carlo simulation of $\boldsymbol{y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with 300 realisations of the parameters:

- $\mathbf{x}_1, \ldots, \mathbf{x}_n \sim N_d(\mathbf{0}, I)$
- $\boldsymbol{\beta}^* \in \mathbb{R}^d$
  - $\beta_1^*, \ldots, \beta^*_{\lfloor d/2 \rfloor} \sim \text{unif}(0,1)$

  - $\beta^*_{\lceil d/2 \rceil}, \ldots, \beta^*_d \sim N(0,1)$
- $\boldsymbol{\beta} = \boldsymbol{\beta}^* ||\boldsymbol{\beta}^*||_2 \quad \text{s.t.} \quad \tau^2 = 1$
- $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, I) \quad \text{s.t.} \quad \sigma^2 = 1$

We compute and plot (see figure 1) the mean values of plug-in estimator for σ^2 as well as their associated 95% confidence intervals for $n = 10000$ and $d = 10 - 500$.

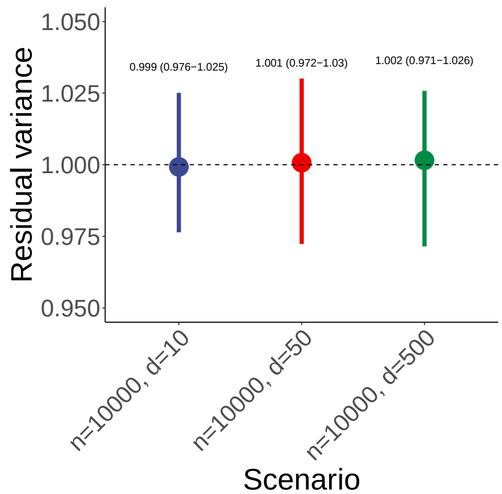

Figure 1: Plug-in estimator for σ^2 in scenarios with $d/n \to 0$

Similarly, we compute and plot (see figure 2) the mean values of plug-in estimator for τ^2 as well as their associated 95% confidence intervals for $n = 10000$ and $d = 10 - 500$.

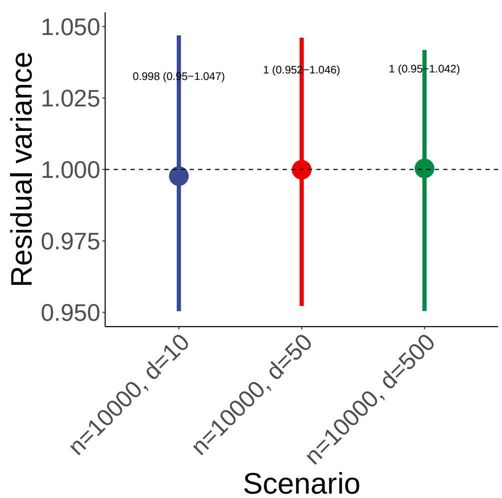

Figure 2: Plug-in estimator for τ^2 in scenarios with $d/n \to 0$

As expected, the estimators for both the noise variance and for the signal strength perform very well. The

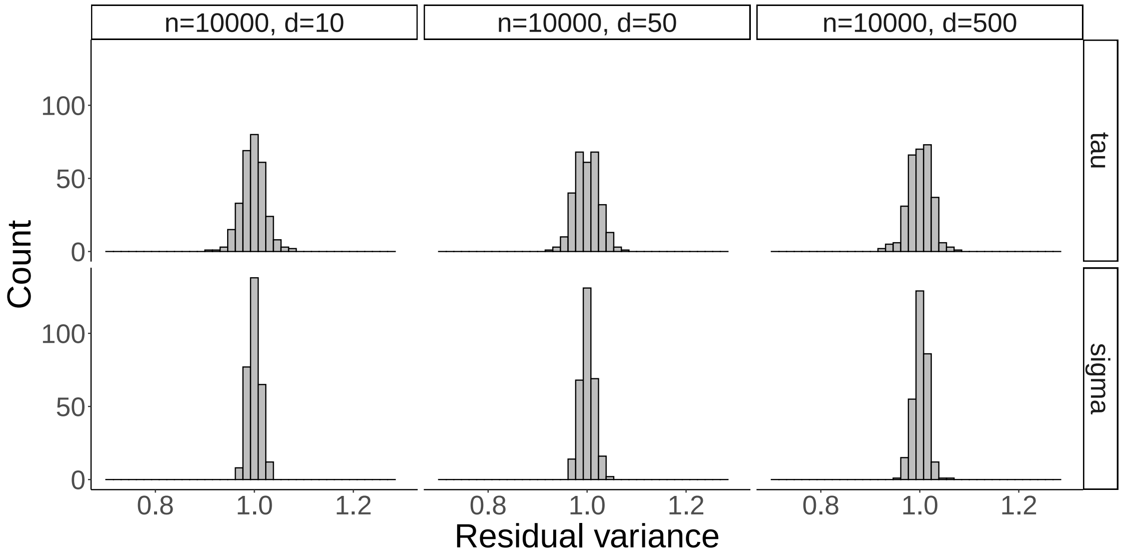

Figure 3: Histograms of plug-in estimator (OLS) for σ^2 and τ^2 in scenarios with $d/n \to 0$

histograms of the plug-in estimator (OLS) for σ^2 and τ^2 are shown in figure 3. We note observe again the good performance of the estimator in the $d/n \to 0$ regime.

3 Regime 2, $\frac{d}{n} \to \rho < 1$

Consider now the case where $\frac{d}{n} \to \rho \leq 1$. We will now consider two sub-scenarios, when Σ is known and when Σ is unknown but estimable. In both cases for simplicity we will assume that the true underlying Σ is the identity matrix.

3.1 Σ is known

We assume that Σ is known. Proceeding in a similar fashion to regime 1 we obtain:

$$\mathbb{E}[\widehat{\sigma}^2_{MLE}] = \frac{\sigma^2}{n}(n-d) = (1-\rho)\sigma^2, \quad \text{which is asymptotically biased}$$

$$\mathbb{V}ar[\widehat{\sigma}^2_{MLE}] = 2\sigma^4 \frac{1-\rho}{n} \to 0, \quad \text{as } n \to \infty$$

We now highlight the work of [2], where he proposes the following unbiased method of moments estimators for the noise variance and signal strength:

$$\hat{\sigma}^2 = \frac{d+n+1}{n(n+1)}||\boldsymbol{y}||_2^2 - \frac{1}{n(n+1)}||\mathbb{X}^T\boldsymbol{y}||_2^2$$

$$\hat{\tau}^2 = -\frac{d}{n(n+1)}||\boldsymbol{y}||_2^2 + \frac{1}{n(n+1)}||X^T\boldsymbol{y}||_2^2$$

We refer the interested reader to the work of Dicker [3, 2] for a full derivation of these results however the key observation is that $||\boldsymbol{y}||_2^2 \sim (\sigma^2+\tau^2)\chi_n^2$ which implies

$$\mathbb{E}\Big(\frac{1}{n}||\boldsymbol{y}||_2^2\Big) = \tau^2 + \sigma^2$$
$$\mathbb{E}\Big(\frac{1}{n^2}||\mathbb{X}^T\boldsymbol{y}||_2^2\Big) = \frac{d+n+1}{n}\tau^2 + \frac{d}{n}\sigma^2$$

The above identities can be written as linear combinations of τ^2 and σ^2. The unbiased estimators of τ^2 and σ^2 may be found by taking linear combinations of $||\boldsymbol{y}||_2^2/n$ and $||\mathbb{X}^T\boldsymbol{y}||_2^2/n^2$.

In the same work Dicker also shows that these estimators are also consistent by proving that

$$\mathbb{V}ar\Big(\sigma^2\Big) = \frac{2}{n}\left(\frac{d}{n}(\tau^2+\sigma^2)^2 + \tau^4 + \sigma^4\right)\left(1 + O\left(\frac{1}{n}\right)\right) \to 0$$
$$\mathbb{V}ar\Big(\tau^2\Big) = \frac{2}{n}\left(\left(1+\frac{d}{n}\right)(\tau^2+\sigma^2)^2 + 3\tau^4 - \sigma^4\right)\left(1 + O\left(\frac{1}{n}\right)\right) \to 0$$

We now proceed to compare Dicker's estimators to the plug-in estimator by numerical simulations.

3.1.1 Numerical simulation

The Monte Carlo simulations are run in a similar manner to the ones described in regime 1. We compute and plot the mean values of plug-in and Dicker estimators for σ^2 (figure 4) and τ^2 (figure 5) as well as their associated 95% confidence intervals for $n = 1000$ and $d = 10 - 500$. Figure 6 shows their corresponding histograms and we can conclude that there is nothing apparently wrong with using the plug-in estimators, which even shows tighter 95% confidence intervals compared to the Dicker estimators. Nevertheless, both type of estimators perform well.

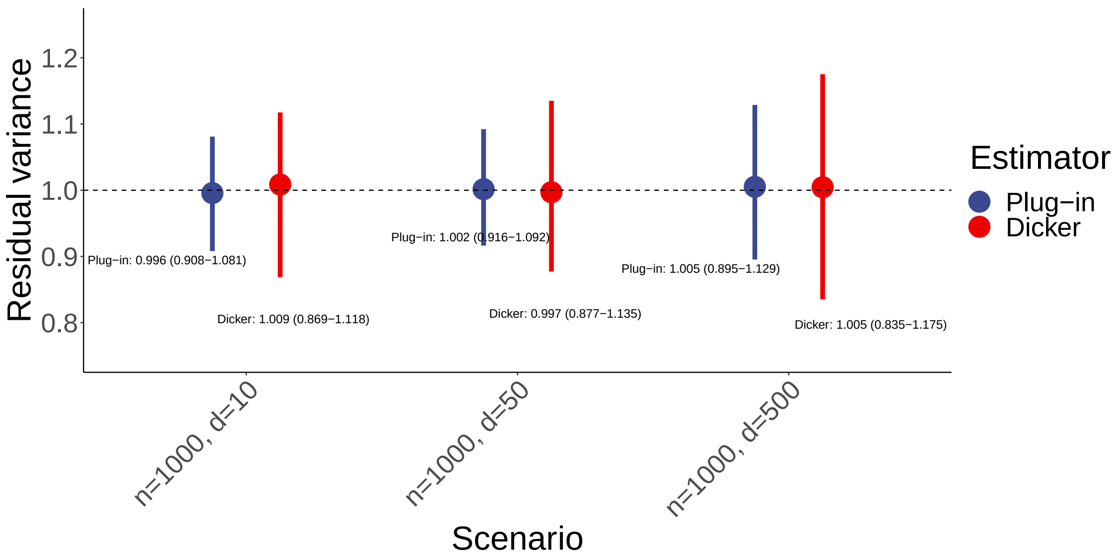

Figure 4: Plug-in and Dicker estimators for σ^2 in scenarios with $d/n \to \rho \leq 1$

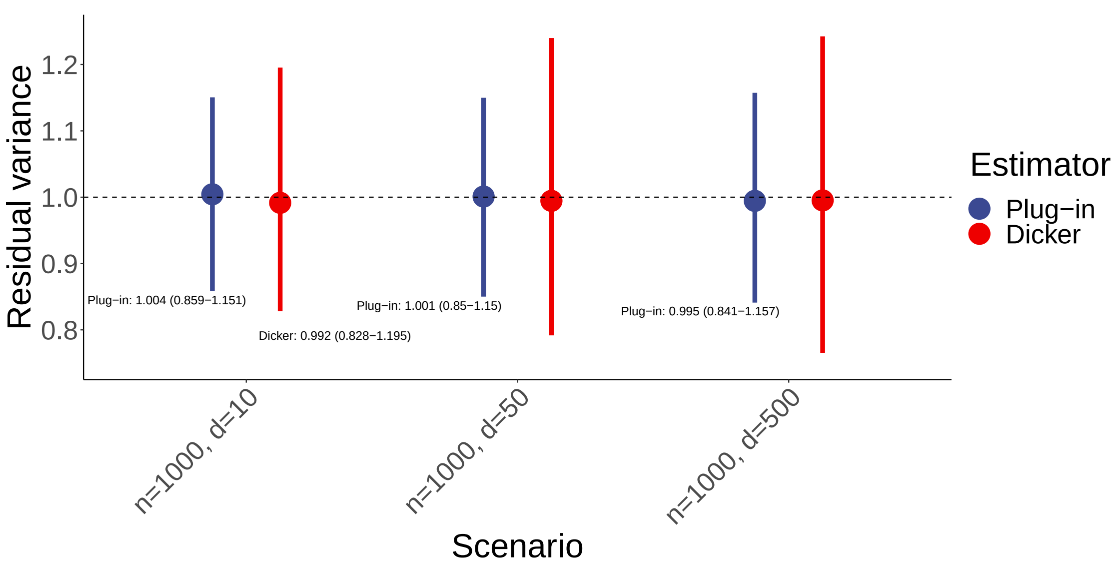

Figure 5: Plug-in and Dicker estimators for τ^2 in scenarios with $d/n \to \rho \leq 1$

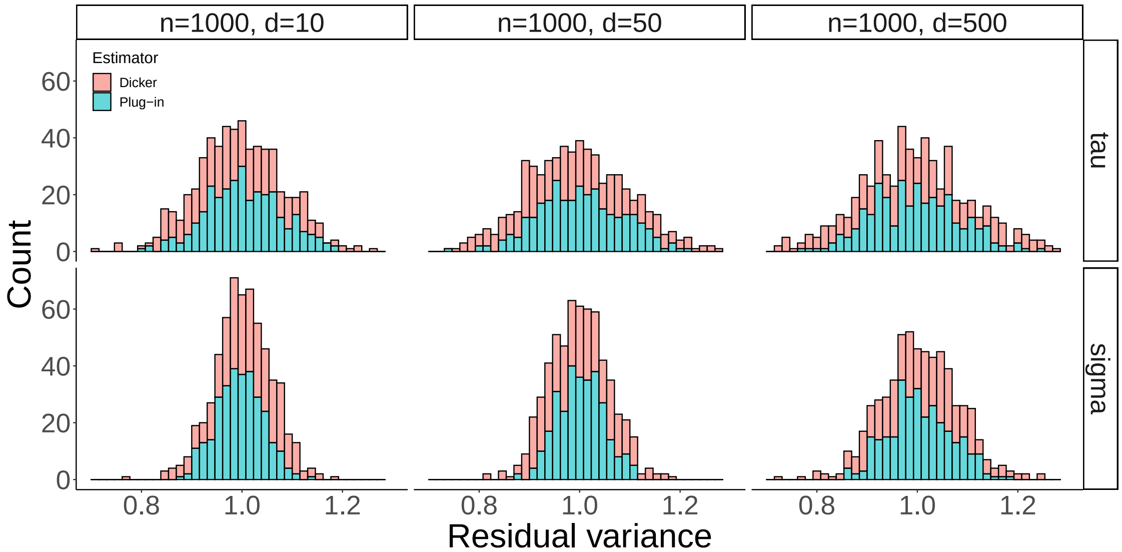

Figure 6: Histograms of plug-in and Dicker estimators for σ^2 and τ^2 in scenarios with $d/n \to \rho \leq 1$ (known Σ)

3.2 Σ is unknown but estimable

We assume that Σ is unknown but estimable. In this scenario we note that both the plug-in estimator and the Dicker estimator have issues, however of different magnitudes:

- Plug-in estimator can no longer be used
- Dicker estimator must be modified

We take the unknown-Σ analogs of known-Σ Dicker estimators $\hat{\sigma}^2$ and $\hat{\tau}^2$. This is done as follows (using some positive definite estimator, $\hat{\Sigma}$):

$$\hat{\sigma}^2(\hat{\Sigma}) = \frac{d+n+1}{n(n+1)}||\boldsymbol{y}||_2^2 - \frac{1}{n(n+1)}||\hat{\Sigma}^{-1/2}\mathbb{X}^T\boldsymbol{y}||_2^2$$
$$\hat{\tau}^2(\hat{\Sigma}) = -\frac{d}{n(n+1)}||\boldsymbol{y}||_2^2 + \frac{1}{n(n+1)}||\hat{\Sigma}^{-1/2}\mathbb{X}^T\boldsymbol{y}||_2^2$$

We will consider two options for estimating Σ:

1. Empiric Σ: $\hat{\Sigma} = \dfrac{1}{n}X^TX$

2. Banded [4] Σ: $\tilde{\Sigma}_{k,p} = B_k(\widehat{\Sigma_p})$, where $B_k(M) = [m_{ij}\mathbb{1}(|i-j| \leq k)]$

$$\begin{bmatrix}
a_{11} & a_{12} & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\
a_{21} & a_{22} & a_{23} & \ddots & & & & \vdots \\
0 & a_{32} & a_{33} & a_{34} & \ddots & & & \vdots \\
\vdots & \ddots & \ddots & \ddots & \ddots & \ddots & & \vdots \\
\vdots & & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\
\vdots & & & \ddots & a_{n-2n-3} & a_{n-2n-2} & a_{n-2n-1} & 0 \\
\vdots & & & & \ddots & a_{n-1n-2} & a_{n-1n-1} & a_{n-1n} \\
0 & \cdots & \cdots & \cdots & \cdots & 0 & a_{nn-1} & a_{nn}
\end{bmatrix}$$

3.2.1 Numerical simulation

Once again, we employ Monte Carlo simulations to evaluate the performance of the empiric and banded Dicker estimators (we employed 5 bands) for Σ, by comparing the estimated values of σ^2 and τ^2. We compute and plot the mean values of these estimators for σ^2 (figure 7) and τ^2 (figure 8) as well as their associated 95% confidence intervals for $n = 1000$ and $d = 10 - 500$. Figure 9 shows their corresponding histograms. We see that as $\rho = d/n$ approaches 1 from below, the banded estimators perform well, while the empiric ones are no longer appropriate.

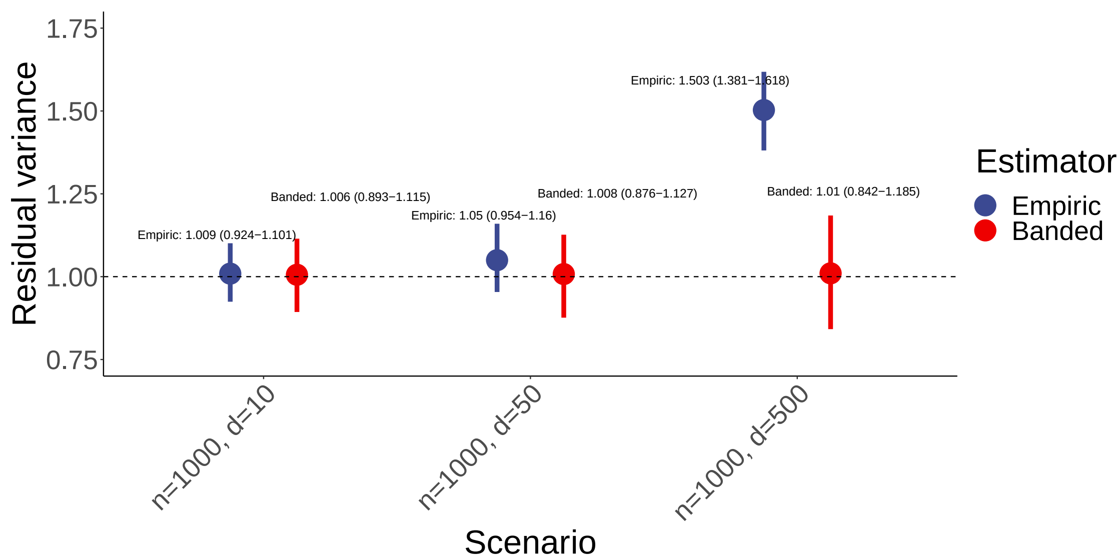

Figure 7: Dicker estimator for σ^2 using empiric and banded estimators of Σ, in scenarios with $d/n \to \rho \leq 1$ (unknown Σ)

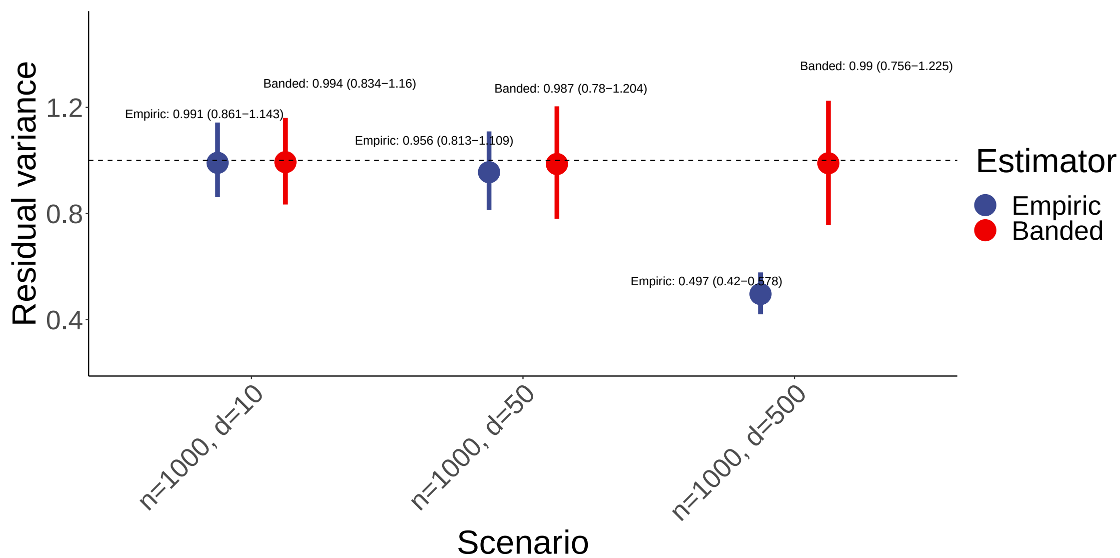

Figure 8: Dicker estimator for τ^2 using empiric and banded estimators of Σ, in scenarios with $d/n \to \rho \leq 1$ (unknown Σ)

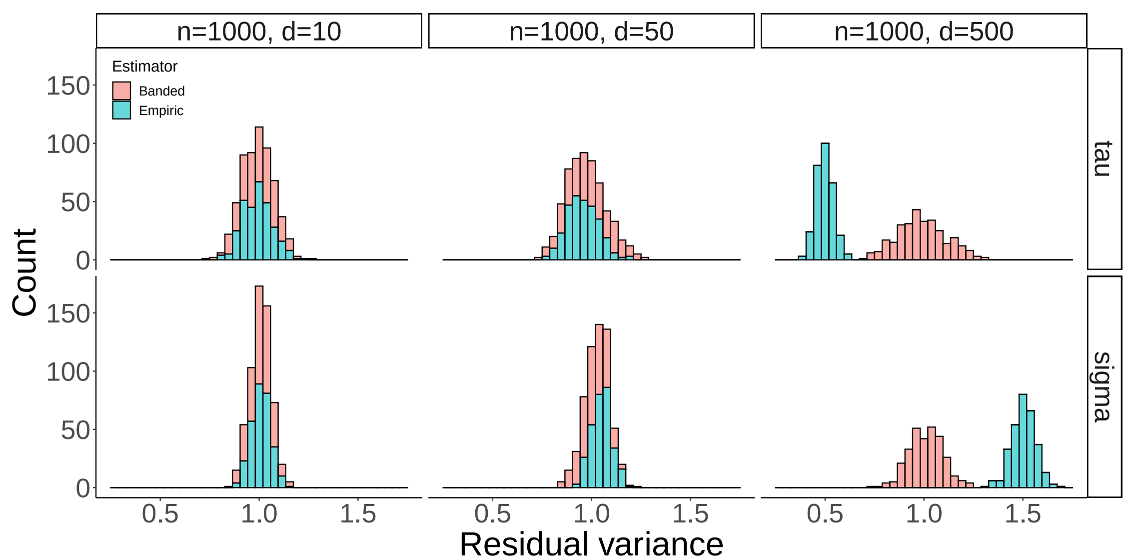

Figure 9: Histograms of Dicker estimator for σ^2 and τ^2 using empiric and banded estimators of Σ, in scenarios with $d/n \to \rho \leq 1$ (unknown Σ)

4 Regime 3, $\frac{d}{n} \to \rho \geq 0$

Consider now our last case of interest where $\frac{d}{n} \to \rho \geq 1$. Here we have $\mathcal{C}(\mathbb{X}) = \mathbb{R}^n$, which implies that $\boldsymbol{\epsilon}^T P_{\mathcal{C}(\mathbb{X})^\perp} \boldsymbol{\epsilon} = 0$. Hence we obtain that $\widehat{\sigma}^2_{MLE} = 0$, which represents the linear model exactly interpolating the data points, not an useful model. Also we cannot use the plug-in estimator as $X^T X$ is not invertible.

4.1 Σ is known

Assume that Σ is known. In order to proceed we introduce another possible estimator for σ^2 and τ^2 called EigenPrism [5]. This estimator is unbiased for σ^2 and τ^2, which are asymptotically normally distributed, and has an estimable tight bound on the variance. There is no need for knowledge of the noise-level, σ^2, or any assumption on the structure of the coefficient vector $\boldsymbol{\beta}$ (e.g. sparsity).

We refer the interested reader to the work on Janson, Barber, and Candès [5] for a full derivation of the EigenPrism estimator, as we provide here only a sketch.

Consider the SVD of $\mathbb{X}$ given by $\mathbb{X} = UDV^T$. Define $\boldsymbol{z} = U^T\mathbf{y}$ and denote the diagonal vector of D by $\boldsymbol{\delta}$. The singular values in D are arranged along the diagonal in decreasing order, such that $\delta_1 \geq \delta_2 \geq \ldots \geq \delta_n \geq 0$. Hence

$$\boldsymbol{z} = D(V^T\boldsymbol{\beta}) + U^T\boldsymbol{\epsilon}$$

One can then show that

$$\mathbb{E}\left(z_i|\boldsymbol{\delta}\right) = \frac{\delta_i^2\tau^2}{d} + \sigma^2$$

Let $\lambda_i = \delta_i^2/d$ be the eigenvalues of $\mathbb{X}\mathbb{X}^T/d$ and let $\mathbf{w} \in \mathbb{R}^n$ be a vector of weights (which need not be non-negative). One can then see that

$$\begin{aligned}
\mathbb{E}\Big(\sum_{i=1}^n w_i z_i^2 \big| \boldsymbol{\delta}\Big) &= \sum_{i=1}^n w_i \Big(\lambda_i\tau^2 + \sigma^2\Big) \\
&= \tau^2 \sum_{i=1}^n w_i\lambda_i + \sigma^2 \sum_{i=1}^n w_i
\end{aligned}$$

Unbiased estimator for τ^2, when constraining $\sum_{i=1}^n w_i = 0$ and $\sum_{i=1}^n w_i\lambda_i = 1$

$$\begin{aligned}
\mathbb{V}ar\Big(\sum_{i=1}^n w_i z_i^2 \big| \boldsymbol{\delta}\Big) &= \tau^2 \sum_{i=1}^n w_i\lambda_i + \sigma^2 \sum_{i=1}^n w_i \\
&\leq 2(\tau^2+\sigma^2)^2 \cdot \max\Big(\sum_{i=1}^n w_i^2, \sum_{i=1}^n (w_i\lambda_i)^2\Big)
\end{aligned}$$

Now define $\mathcal{P}_1 = \operatorname{argmin}_{w\in\mathbb{R}^n} \max\Big(\sum_{i=1}^n w_i^2, \sum_{i=1}^n (w_i\lambda_i)^2\Big)$, s.t. $\sum_{i=1}^n w_i = 0$, $\sum_{i=1}^n w_i\lambda_i = 1$, and with solution w^*.

One then can obtain the EigenPrism estimator as characterised by:

$$\mathbb{E}\left(\sum_{i=1}^{n} w_i^* z_i^2 | \boldsymbol{\delta}\right) = \hat{\tau}^2$$

$$\mathrm{SD}\left(\sum_{i=1}^{n} w_i^* z_i^2 | \boldsymbol{\delta}\right) \lessapprox \sqrt{2\mathrm{val}(\mathcal{P}_1)}\frac{||\mathbf{y}||_2^2}{n}$$

where $\tau^2 + \sigma^2$ was approximated by its estimator $||\mathbf{y}||_2^2/n$.

4.1.1 Numerical simulation

As before, we employ Monte Carlo simulations to evaluate the performance of the Dicker and EigenPrism estimators for σ^2 and τ^2. We compute and plot the mean values of these estimators for σ^2 (figure 10) and τ^2 (figure 11) as well as their associated 95% confidence intervals for $n = 1000$ and $d = 1200$. We see that both estimators perform well.

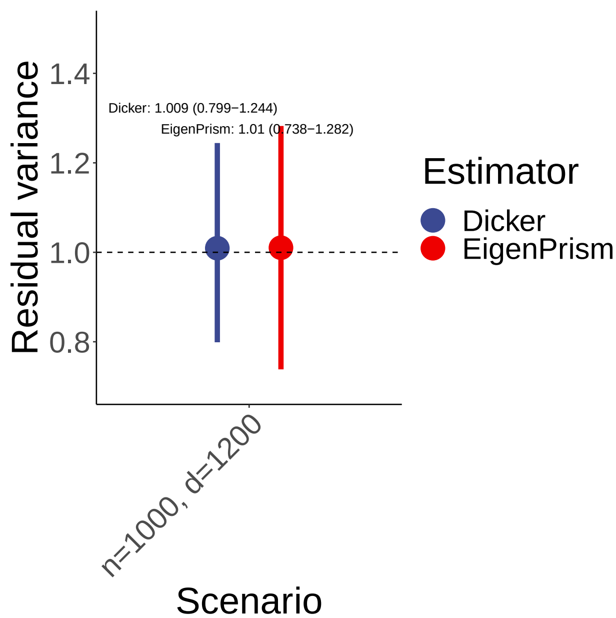

Figure 10: Dicker and EigenPrism estimators for σ^2 in scenarios with $d/n \to \rho > 1$ (known Σ)

4.2 Σ is unknown but estimable

Finally, we briefly investigate numerically the case when Σ is unknown but estimable. We note that the empiric variance-covariance matrix, $\mathbb{X}^T\mathbb{X}/n$, no longer norm-consistent for Σ. Also, it is generally not possible to find norm-consistent estimator for Σ, without further information/assumptions. However, for the sake of comparison and academic inquiry we will define an empiric estimator using the Moore-Penrose generalized inverse of Σ.

Assuming some banded structure for the underlying Σ, we employ a banded Dicker style estimator for σ^2 and τ^2 and perform Monte Carlo simulation. We compute and plot the mean values of empiric and Dicker

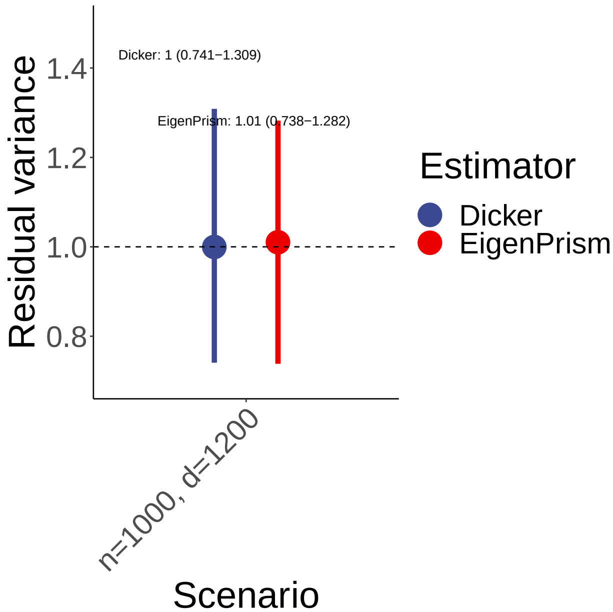

Figure 11: Dicker and EigenPrism estimators for τ^2 in scenarios with $d/n \to \rho > 1$ (known Σ)

estimators for σ^2 (figure 12) and τ^2 (figure 13) as well as their associated 95% confidence intervals for $n = 1000$ and $d = 1200$. We see while the Dicker estimator perform well, the empirical one clearly is not applicable in this regime

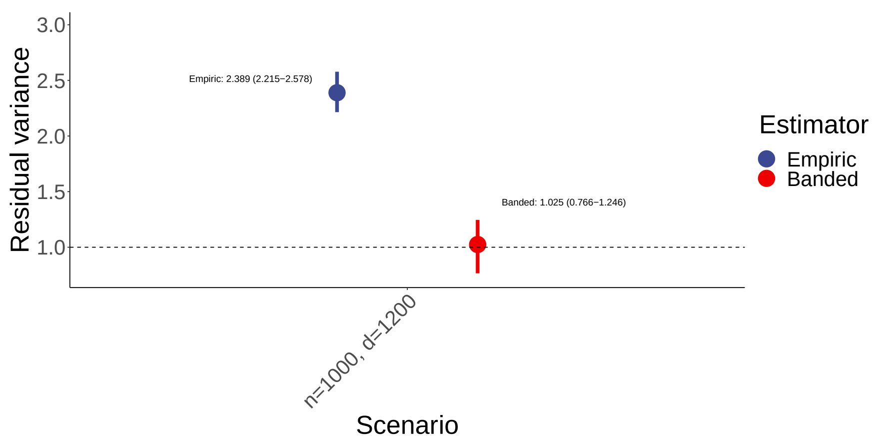

Figure 12: Empiric and Dicker estimators for σ^2 in scenarios with $d/n \to \rho > 1$ (unknown Σ)

5 Conclusions

Throughout this report, we investigated various estimators for the noise variance and signal strength. We observed how as the ratio between the number of dimensions and the number of samples, ρ, increases from zero and towards one, the plug-in estimators actually performs rather well, as long as the structure of Σ is

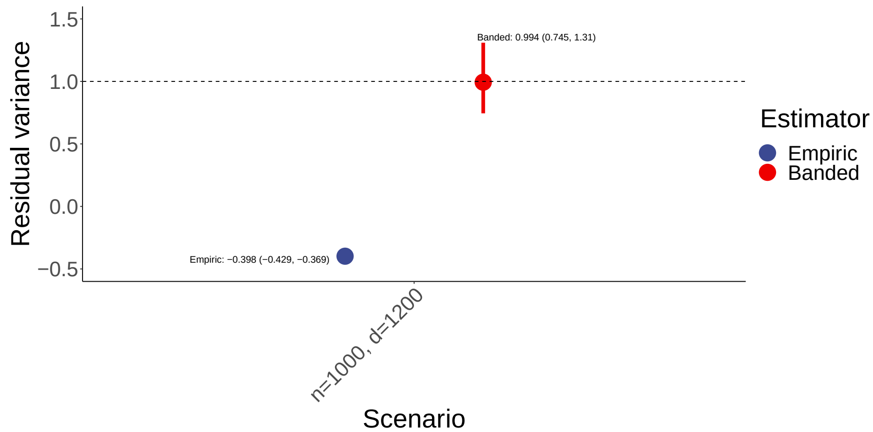

Figure 13: Empiric and Dicker estimators for τ^2 in scenarios with $d/n \to \rho > 1$ (unknown Σ)

known. However, as ρ approaches one, if Σ is unknown then the empiric estimator no longer is appropriate and a Dicker estimator is preferred. Furthermore, for $\rho > 1$, empiric estimators perform badly regardless of Σ, and Dicker and EigenPrism estimators show their strength. If interested in the simulations, please visit: `github.com/arinmadenci/bst-235-project`

We would like to thank Professor Mukherjee for an insightful and entertaining course, and for giving us the chance the expand beyond the lecture material into the fun (but hard) field of high-dimensional statistics. We hope you enjoyed reading our report. Happy Holidays!

References

[1] Morris L. Eaton and Michael D. Perlman. The non-singularity of generalized sample covariance matrices. *Ann. Statist.*, 1(4), 1973.

[2] Lee H. Dicker. Variance estimation in high-dimensional linear models. *Biometrika*, 101(2), 2014.

[3] Lee H. Dicker. Optimal equivariant prediction for high-dimensional linear models with arbitrary predictor covariance. *Electron. J. Statist.*, 7, 2013.

[4] Peter J. Bickel and Elizaveta Levina. Regularized estimation of large covariance matrices. *The Annals of Statistics*, 36(1):199–227, February 2008.

[5] Lucas Janson, Rina Foygel Barber, and Emmanuel Candès. Eigenprism: inference for high dimensional signal-to-noise ratios. *J. R. Statist. Soc. B*, 79, 2016.