Functional Estimation in High Dimensional Problems

Arin Madenci and Tudor Cristea-Platon

BST235

December 2019

Introduction

To understand the behavior of models and performance limits of model-fitting procedures, **residual variance** and **proportion of explained variance** are critical estimands.

Well-known applications, such as:

- Information about scale of an estimator's risk under ℓ^2 loss
- Computation of model selection statistics
- Regression diagnostics (e.g., goodness-of-fit testing)
- Signal-to-noise ratio

Motivation

Estimators of residual variance and proportion of explained variance may or may not be reliable in high-dimensional conditions.

We will evaluate several estimators:

- "Plug-in" estimator (OLS)
- Dicker 2013
- "EigenPrism," Janson, Barber, and Candes 2017

Motivation

After briefly reviewing notation, assumptions, and definitions, we will compare their performance under a variety of conditions of d (number of parameters) and n (number of observations):

- $\frac{d}{n} \to 0$
- $\frac{d}{n} \to \rho \leq 1$, with Σ known
- $\frac{d}{n} \to \rho \leq 1$, with Σ unknown
- $\frac{d}{n} \to \rho > 1$, with Σ known
- $\frac{d}{n} \to \rho > 1$, with Σ unknown

Notation

- Linear model $y_i = x_i^T \beta + \epsilon_i,\ i = 1, \ldots, n$
- $y = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$ denotes the n-dimensional vector of observed outcomes
- $X = (x_1, \ldots, x_n)^T \in \mathbb{R}^n$ denotes the $n \times d$ matrix of observed predictors, where $x_1 = (x_{11}, \ldots, x_{1d})^T, \ldots, x_n = (x_{n1}, \ldots, x_{nd})^T \in \mathbb{R}^d$ are d-dimensional predictors
- $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T \in \mathbb{R}^n$ is a vector of unobserved i.i.d. errors with $\mathbb{E}(\epsilon_i) = 0$ and $\mathrm{var}(\epsilon_i) = \sigma^2 > 0$
- $\beta = (\beta_1, \ldots \beta_d)^T \in \mathbb{R}^d$ is an unknown d-dimensional parameter

Assumptions

- Random predictors x_i with mean $\mathbb{E}(x_i) = 0$ and $d \times d$ positive definite covariance matrix $\mathrm{cov}(x_i) = \Sigma$
- $x_i \perp\!\!\!\perp \epsilon$
- $\epsilon_1, \ldots, \epsilon_n \sim N(0, \sigma^2)$
- $x_1, \ldots, x_n \sim N(0, \Sigma)$
- Note: $\mathbb{E}(x_i) = 0$, without loss of generality; for cases in which $\mathbb{E}(x_i) \neq 0$, include intercept term, center data, and replace n with $n - 1$.

Definitions: Signal and Noise

- τ^2: ℓ^2-signal strength
  - Let $\tau^2 = \beta^T \Sigma \beta = ||\Sigma^{1/2}\beta||^2$
  - Where $||\cdot||$ is the ℓ^2-norm
- σ^2: residual variance (noise)
  - $\sigma^2 = \mathbb{V}\mathrm{ar}(\epsilon_i) = \mathbb{V}\mathrm{ar}\Big(\mathbb{E}(y_i|x_i)\Big)$

Returning to our question of interest:

$$\boxed{\text{How can we identify effective estimators of these quantities (i.e., } \tau^2 \text{ and } \sigma^2\text{) in high-dimensional linear models with large } d \text{ and } n \text{ ... and especially for } d > n\text{?}}$$

Scenario #1: $d/n \to 0$

For estimation of σ^2:

- $\widehat{\sigma}^2_{MLE} = \frac{1}{n}||P_{\mathcal{C}(\mathbb{X})^\perp}(\mathbf{y})||^2$ is the MLE 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 $\frac{\sigma^2}{n}\left(\frac{\boldsymbol{\epsilon}}{\sigma}\right)^T P_{\mathcal{C}(\mathbb{X})^\perp}\left(\frac{\boldsymbol{\epsilon}}{\sigma}\right)\Big|\,\mathbb{X} \sim \chi^2_{n-\text{rank}(\mathbb{X})}(0)$.
- From previous work[1], we know that for $d \leq n$, $P(\text{rank}(\mathbb{X} = d)) = 1$ as $\mathbf{x_i}$ is continuous.

[1] Eaton ML and Perlman MD. The Non-Singularity of Generalized Sample Covariance Matrices. The Annals of Statistics 1973;1:710–7.

Scenario #1: $d/n \to 0$

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

This implies:

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

$$\mathbb{V}ar(\hat{\sigma}^2_{MLE}) = \frac{\sigma^4}{n^2} 2(n-d) \to 0 \quad \text{as} \quad n \to \infty$$

Result: $\hat{\sigma}^2_{MLE}$ and $\hat{\tau}^2_{MLE}$ (not shown) unbiased for $d/n \to 0$.

Scenario #1: $d/n \to 0$

For MLE estimation of τ^2:

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

Scenario #1: $d/n \to 0$

Bias of $\hat{\tau}^2_{MLE}$:

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

For $\Sigma = I$:

$$= \sigma^2\frac{d}{n(n-d-1)}$$

Numerical simulation: $\sigma^2 = \tau^2 = 1$

- $x_1, \ldots, x_n \in \mathbb{R}^d \sim N(0, I)^\dagger$
- $\beta^* \in \mathbb{R}^d$
  - $\beta^*_1, \ldots, \beta^*_{\lfloor d/2 \rfloor} \sim \mathrm{unif}(0,1)$
  - $\beta^*_{\lceil d/2 \rceil}, \ldots, \beta^*_d \sim N(0,1)$
- $\beta = \beta^* ({\beta^*}^T \beta^*)^{-1/2} \quad \text{s.t.} \quad \tau^2 = 1$
- $Y = X\beta + \epsilon, \quad \epsilon \sim N(0,1) \quad \text{s.t.} \quad \sigma^2 = 1$
- Monte Carlo simulation with 300 replications

† Note: assume $\Sigma = I$ without loss of generality (for $\Sigma \neq I$, replace (X, β) with $(X\Sigma^{-1/2}, \Sigma^{1/2}\beta)$.

Numerical simulation for σ^2: $d/n \to 0$

Plug-in estimator (OLS) performs well, as expected:

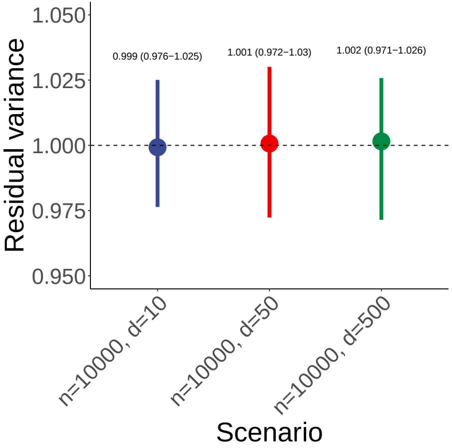


Figure: Plug-in estimator for σ^2 in scenarios with $d/n \to 0$

Numerical simulation for τ^2: $d/n \to 0$

Again, plug-in estimator (OLS) performs well

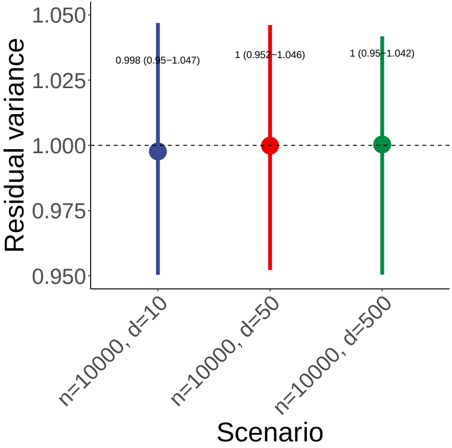


Figure: Plug-in estimator for τ^2 in scenarios with $d/n \to 0$

Numerical simulation: $d/n \to 0$

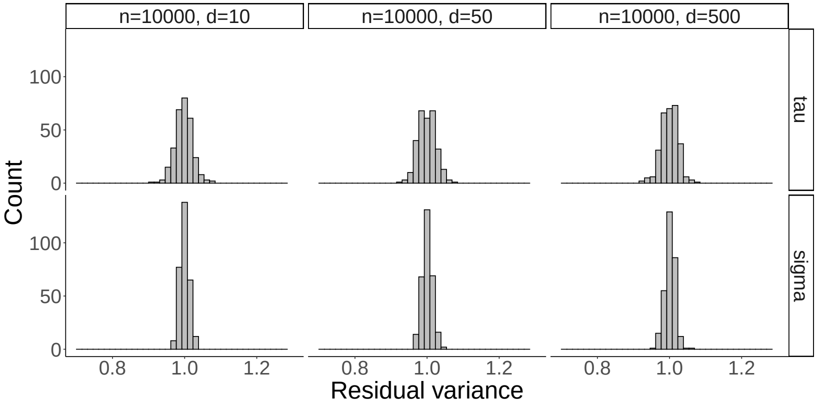


Figure: Histograms of plug-in estimator for σ^2 and τ^2 in scenarios with $d/n \to 0$

Scenario #2: $d/n \to \rho \leq 1$

Plug-in estimator:

$$\mathbb{E}[\widehat{\sigma}^2_{MLE}] = \sigma^2 \frac{d}{n(n-d-1)}$$

Dicker (2014) proposes the following method of moments estimators:

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

Scenario #2: $d/n \to \rho \leq 1$ (known Σ)

Sketch of proof:[2]

$$||y||_2^2 \sim (\sigma^2 + \tau^2)\chi_n^2 \implies \mathbb{E}\left(\frac{1}{n}||y||_2^2\right) = \tau^2 + \sigma^2$$
$$\mathbb{E}\left(\frac{1}{n^2}||X^T y||_2^2\right) = \frac{d+n+1}{n}\tau^2 + \frac{d}{n}\sigma^2$$

- Above identities can be written as linear combinations of τ^2 and σ^2.
- Unbiased estimators of τ^2 and σ^2 may be found by taking linear combinations of $n^{-1}||y||_2^2$ and $n^{-2}||X^T y||_2^2$.
- Unbiased estimators in the case that $\Sigma = I$.

[2] Full reference available in Dicker 2014.

Numerical simulation for σ^2: $d/n \to \rho \leq 1$ (known Σ)

Is it so bad to use our plug-in estimator for scenarios in which $d/n \to \rho \leq 1$ (as opposed to $d/n \to 0$)?

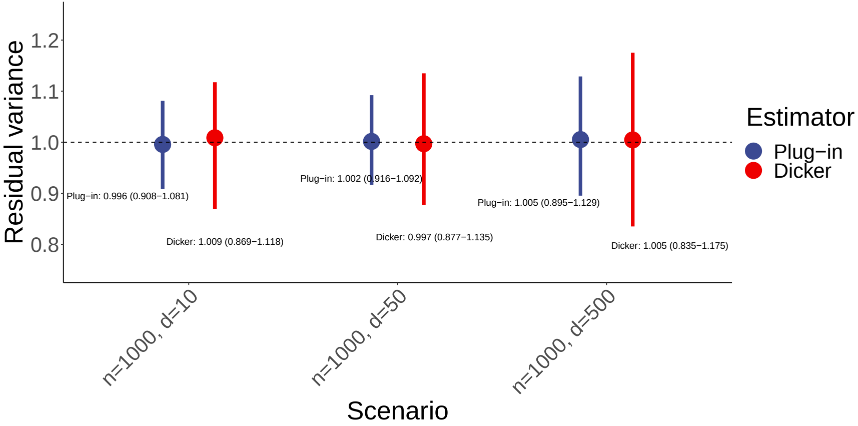


Figure: Plug-in and Dicker estimators for σ^2 in scenarios with $d/n \to \rho \leq 1$

Numerical simulation for τ^2: $d/n \to \rho \le 1$ (known Σ)

Is it so bad to use our plug-in estimator for scenarios in which $d/n \to \rho \le 1$ (as opposed to $d/n \to 0$)?

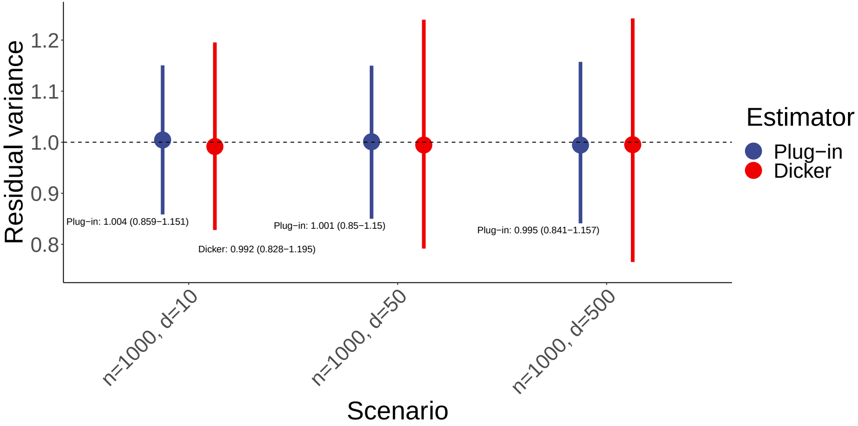


Figure: Plug-in and Dicker estimators for τ^2 in scenarios with $d/n \to \rho \le 1$

Numerical simulation: $d/n \to \rho \leq 1$ (known Σ)

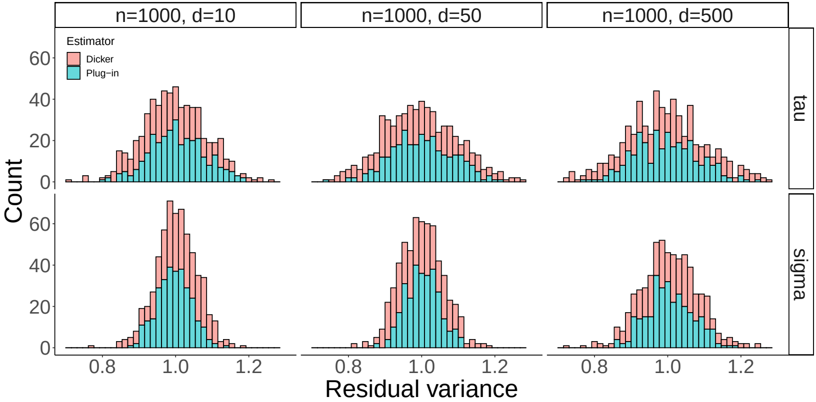


Figure: Histograms of plug-in and Dicker estimators for σ^2 and τ^2 in scenarios with $d/n \to \rho \leq 1$ (known Σ)

Scenario #3: $d/n \to \rho \leq 1$, Σ unknown (estimable)

With unknown Σ:

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

Take the unknown-Σ analogs of known-Σ Dicker estimators $\hat{\sigma}^2$ and $\hat{\tau}^2$ above, as follows (using some positive definite estimator, $\hat{\Sigma}$):

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

Scenario #3: $d/n \to \rho \leq 1$, Σ unknown

We will consider two options for estimating Σ:

1. Empiric Σ: $\hat{\Sigma} = \frac{1}{n} X^T X$

2. Banded Σ: $\tilde{\Sigma}_{k,p} = B_k(\hat{\Sigma}_p)$

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

Where $B_k(M) = [m_{ij} \mathbb{1}(|i-j| \leq k)]$

Numerical simulation for σ^2: $d/n \to \rho \leq 1$, Σ unknown

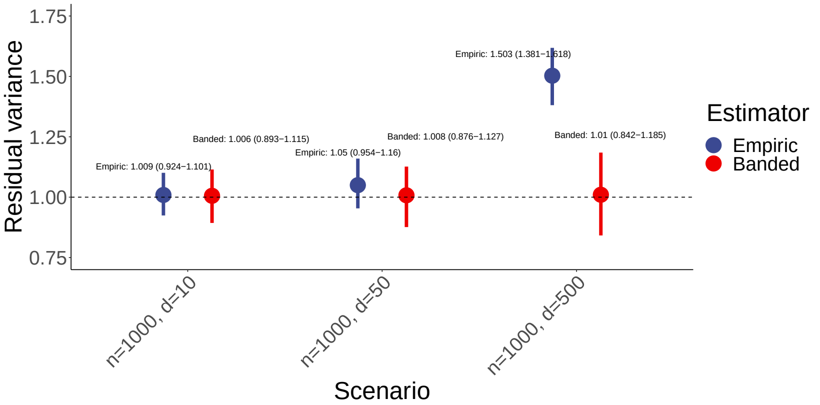


Figure: Numerical simulation for σ^2

Note: # bands, $k = 5$ for simulations.

Numerical simulation for τ^2: $d/n \to \rho \leq 1$, Σ unknown

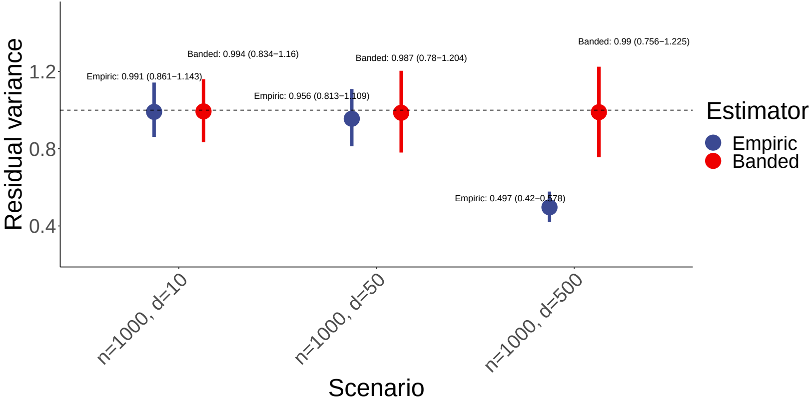


Figure: Numerical simulation for τ^2

Numerical simulation for τ^2: $d/n \to \rho \leq 1$, Σ unknown

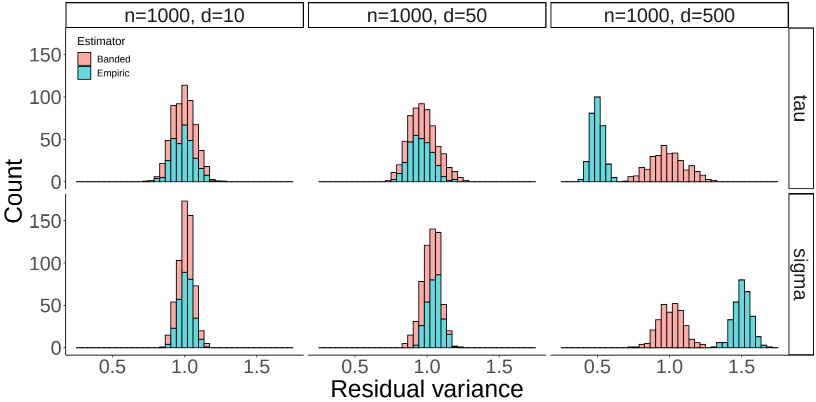


Figure: Histograms of Dicker estimator for σ^2 and τ^2 using empiric and banded estimators of Σ, in scenarios with $d/n \to \rho \leq 1$ (unknown Σ)

$d/n \to \rho > 1$, Σ known

Introduce another possible estimator for σ^2 and τ^2: EigenPrism (Janson, Barber, and Candès):

- Goal to develop estimators unbiased for σ^2 and τ^2, which are asymptotically normally distributed, and with estimable tight bound on variance.
- No need for knowledge of the noise-level (σ^2) or any assumption on the structure of the coefficient vector β (e.g. sparsity).

$d/n \to \rho > 1$, Σ known

Sketch of proof for EigenPrism estimators:

$$\mathbb{E}\Big(\sum_{i=1}^{n} w_i z_i^2 \Big| d\Big) = \sum_{i=1}^{n} w_i \Big(\lambda_i \tau^2 + \sigma^2\Big)$$
$$= \tau^2 \sum_{i=1}^{n} w_i \lambda_i + \sigma^2 \sum_{i=1}^{n} w_i\Big)$$

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

$d/n \to \rho > 1$, Σ known

$$\mathbb{V}ar\left(\sum_{i=1}^{n} w_i z_i^2 \middle| d\right) = \tau^2 \sum_{i=1}^{n} w_i \lambda_i + \sigma^2 \sum_{i=1}^{n} w_i$$
$$\leq 2(\theta^2 + \sigma^2)^2 \cdot \max\left(\sum_{i=1}^{n} w_i^2, \sum_{i=1}^{n} (w_i \lambda_i)^2\right)$$

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

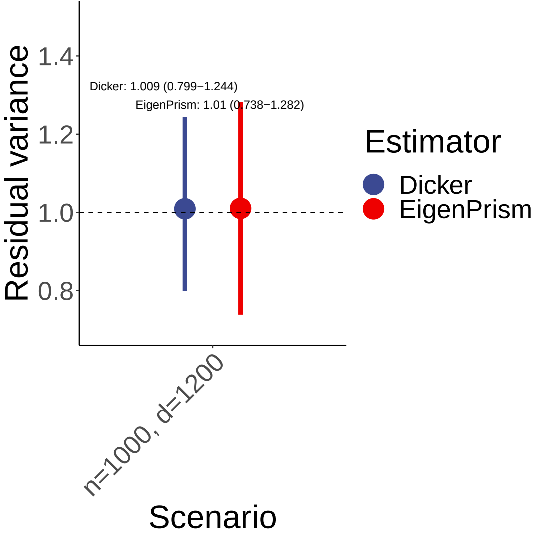
EigenPrism:

- $\mathbb{E}(\sum_{i=1}^n w_i^* z_i^2 | d) = \hat{\tau}^2$
- $\mathrm{SD}(\sum_{i=1}^n w_i^* z_i^2 | d) \lessapprox \sqrt{2\mathrm{val}(\mathcal{P}_1)} \frac{||y||_2^2}{n}$

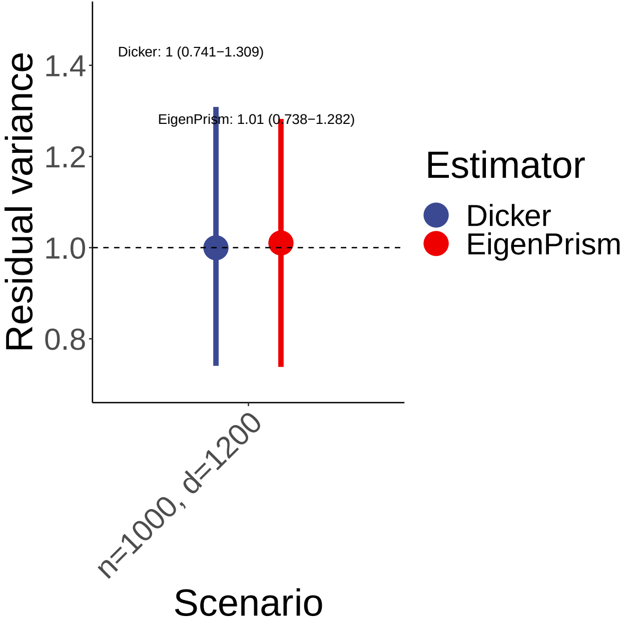
Full proof in reference: Janson L, Barber RF, and Candès 2017

Numerical simulation for σ^2: $d/n \to \rho > 1$, Σ known

With $d > n$, cannot use plug-in estimator ($X^T X$ non-invertible). Dicker estimator performs well.



Numerical simulation for τ^2: $d/n \to \rho > 1$, Σ known



Numerical simulation for τ^2: $d/n \to \rho > 1$, Σ known

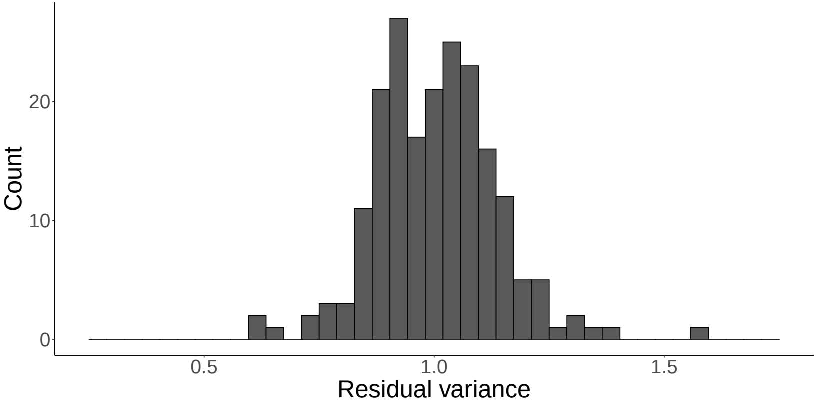


Figure: Histograms of Dicker estimator for σ^2 and τ^2 using empiric and banded estimators of Σ, $d = 1200, n = 1000$ (known Σ)

Numerical simulation for σ^2: $d/n \to \rho > 1$, Σ unknown

- Empiric covariance matrix, $n^{-1}X^TX$ no longer norm-consistent for Σ.
- Generally not possible to find norm-consistent estimator for Σ, without further information/assumptions.

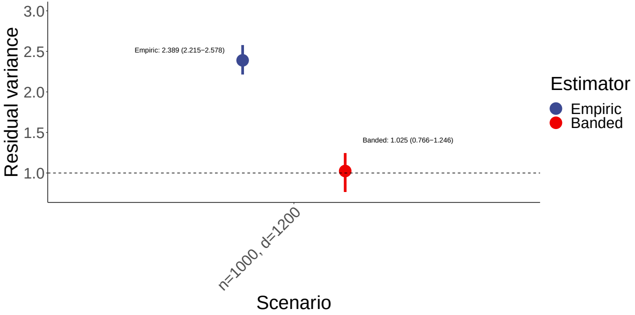


Figure: Numerical simulation for σ^2

Note: Empiric estimator now requires Moore-Penrose generalized inverse of Σ.

Numerical simulation for τ^2: $d/n \to \rho > 1$, Σ unknown

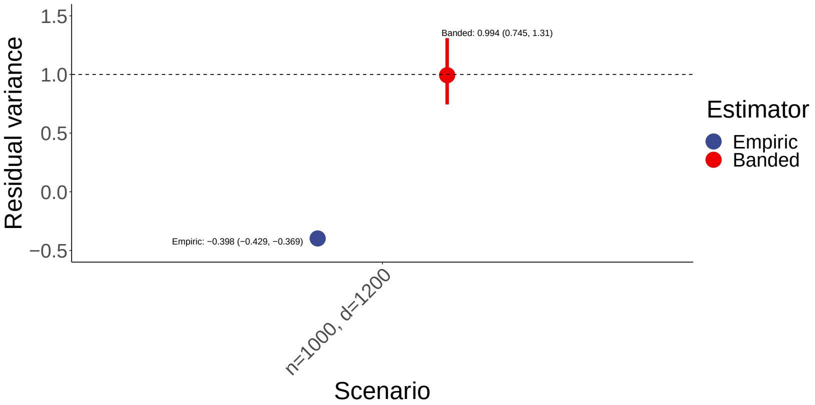


Figure: Numerical simulation for τ^2

References

Eaton ML and Perlman MD. The Non-Singularity of Generalized Sample Covariance Matrices. The Annals of Statistics 1973;1:710–7.

Dicker LH. Optimal equivariant prediction for high-dimensional linear models with arbitrary predictor covariance. Electronic Journal of Statistics 2013;7:1806–34.

Dicker LH. Variance estimation in high-dimensional linear models. Biometrika 2014;101:269–84.

Janson L, Barber RF, and Candès E. EigenPrism: inference for high dimensional signal-to-noise ratios. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 2017;79:1037–65.

Bickel PJ and Levina E. Regularized estimation of large covariance matrices. The Annals of Statistics 2008;36:199–227.