Quantifying Optimism and Model Comparison in Matrix Approximation: a Degrees of Freedom Approach

Luis Sierra Muntané
luis.sierra@mail.utoronto.ca

DoSS Student Research Day 2026
University of Toronto DoSS

24th April 2026

Two main goals

- Analytical, efficient estimation of model optimism for risk estimation.
- Comparison of model fit through a notion of *effective parameters*.

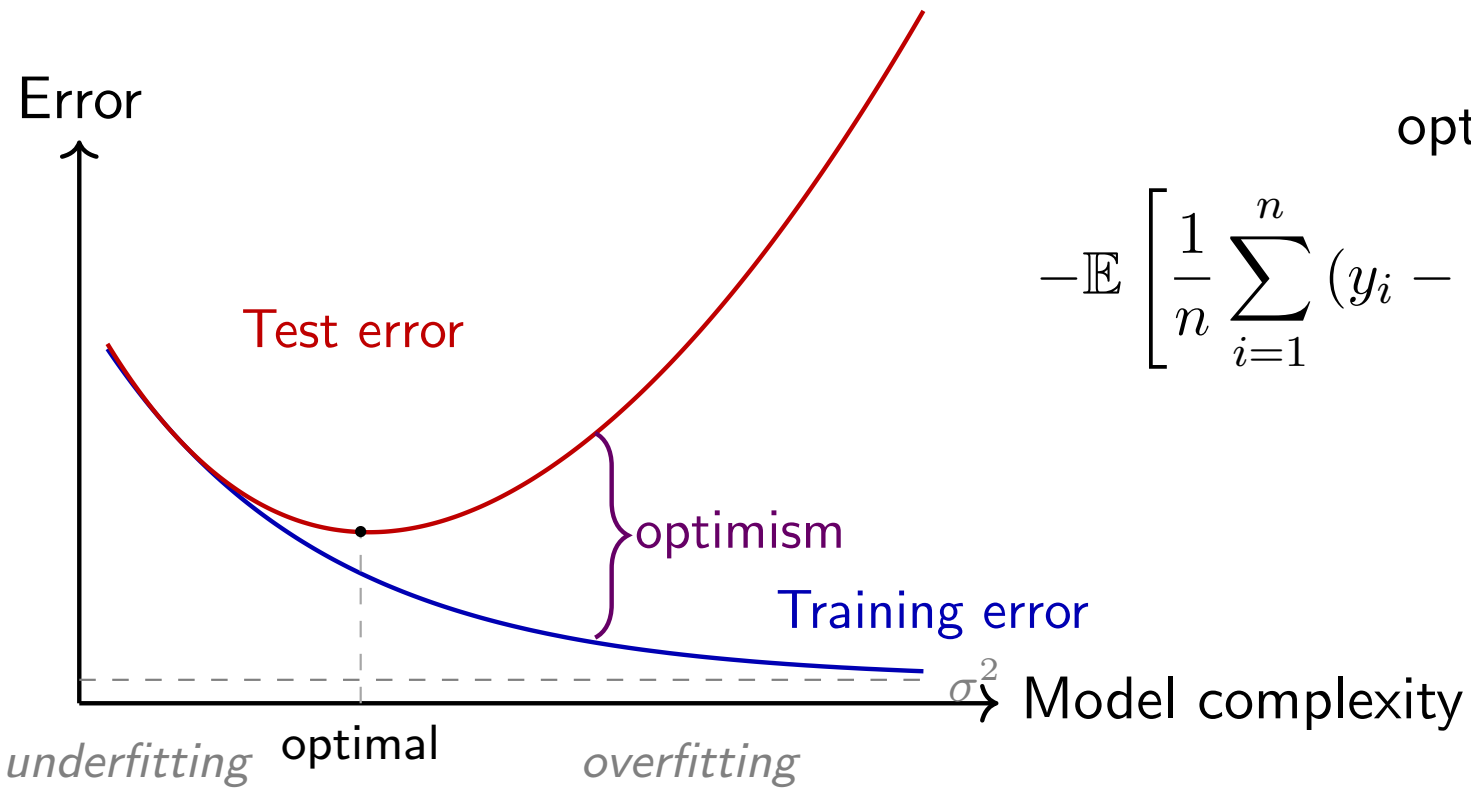
Outline:

Model Optimism

The error of any fitting procedure $\hat{f}$ on our dataset (training error) is a poor proxy for the true error. Consider a prediction setting

$$y_i = f(x_i) + \varepsilon_i,$$

where f is some regression function $f(x) = \mathbb{E}[y_i \mid x_i = x]$ and ε_i are mean-zero iid. errors.



$$\text{opt}(\hat{f}) = \text{err}(\hat{f})$$

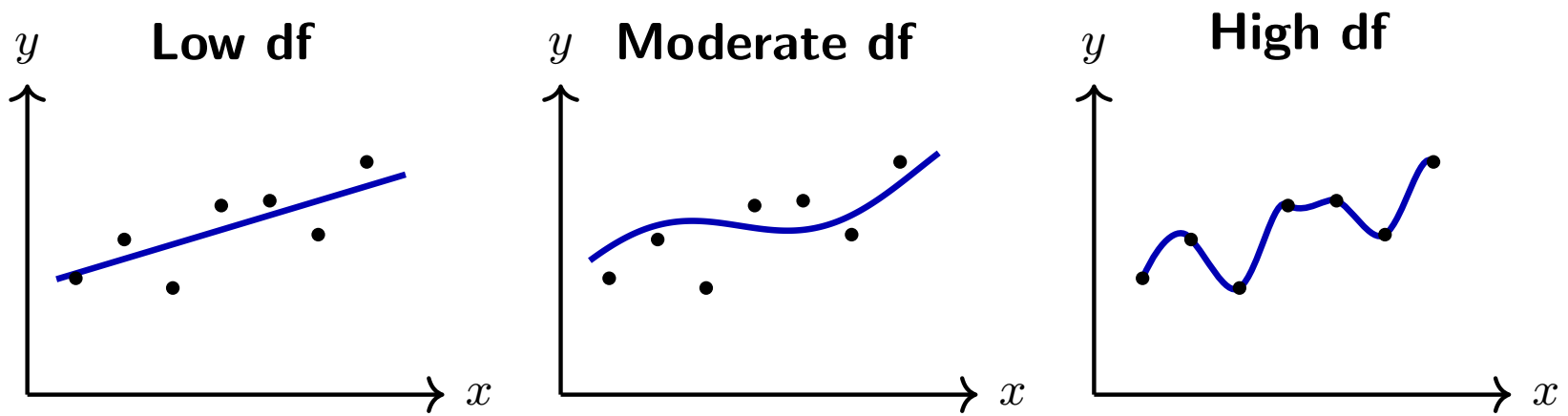
$$-\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2 \mid X\right]$$

Optimism and degrees of freedom

Theorem ((Efron, 1986, 2004))

For random error with variance σ^2, the model optimism is given exactly by

$$\text{opt}(\hat{f}) = \frac{2}{n}\sum_{i=1}^{n}\text{Cov}(\hat{f}(x_i), y_i) =: \frac{2\sigma^2}{n}\text{df}(\hat{f}).$$



df $= 2$ (linear fit) df ≈ 4 (smoother fit) df $= n = 7$ (interpolation)

Unbiased Risk Estimation

We can decompose the Risk $\mathbb{E}\|\hat{f}(x) - f(x)\|_2^2$ as

$$\mathsf{R}(\hat{f}) = \|y - \hat{f}(x_i)\|^2 - n\sigma^2 + 2\sigma^2 \mathsf{df}(\hat{f}), \quad \mathsf{df}(\hat{f}) = \frac{1}{\sigma^2} \sum_{i=1}^{n} \mathsf{Cov}(\hat{f}(x_i), y_i)$$

Proposition ((Stein, 1981; Tibshirani and Wasserman, 2015))

For Gaussian noise, an unbiased estimate of the Risk of an almost differentiable estimator $\hat{f}$ is

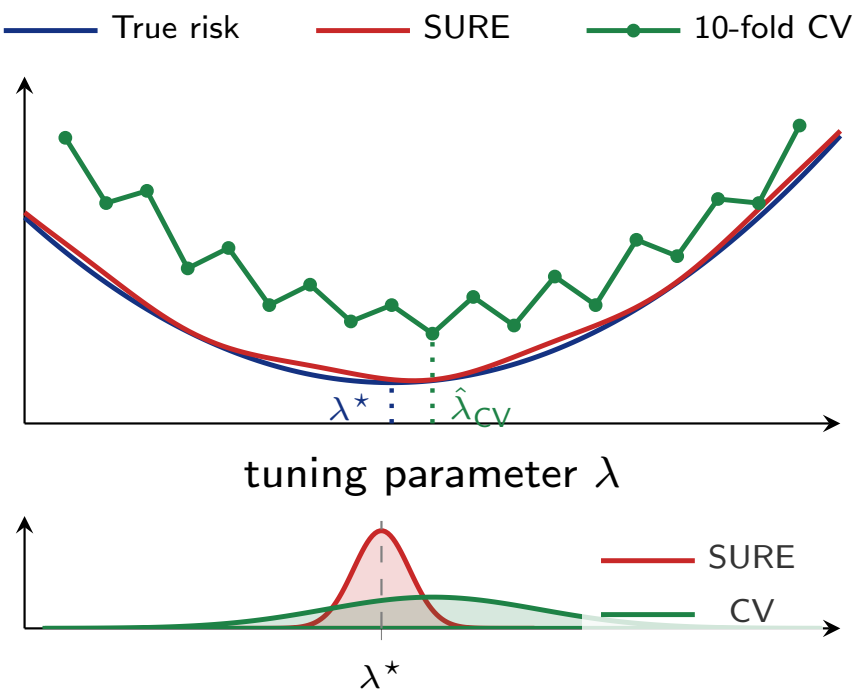
$$\mathsf{SURE}(\hat{f}) = \|y - \hat{f}(x_i)\|^2 - n\sigma^2 + 2\sigma^2 \widehat{\mathsf{df}}(\hat{f}), \quad \widehat{\mathsf{df}}(\hat{f}) = \mathsf{tr}\left[\nabla \hat{f}.\right]$$

> *Beyond Gaussian noise, there has been some exploration into Exponential family noise by Eldar (2008) and using Stein's Method to bound the dissimilarity from a Gaussian distribution.*

Why SURE Beats Cross-Validation for Risk Estimation

$$\widehat{\mathrm{R}}(\hat{f}) = \|y - \hat{f}(x_i)\|_2^2 - n\sigma^2 + 2\sigma^2 \widehat{\mathrm{df}}(\hat{f}), \quad \widehat{\mathrm{df}}(\hat{f}) = \sum_{i=1}^{n} \frac{\partial \hat{f}(x_i)}{\partial y_i}$$

- **Unbiased** for the true risk under the *Gaussian model*; K-fold CV is biased upward since each fold trains on only $(K-1)/K$ of the data.
- **Single fit.** CV refits K times (n for LOOCV) computationally expensive for MCMC, deep nets, etc.
- **Deterministic and lower variance.**
- **Curve shape.** (Dramatized in diagram) CV curves can have "incorrect" shapes; minimizers are unstable (Bates et al., 2023).

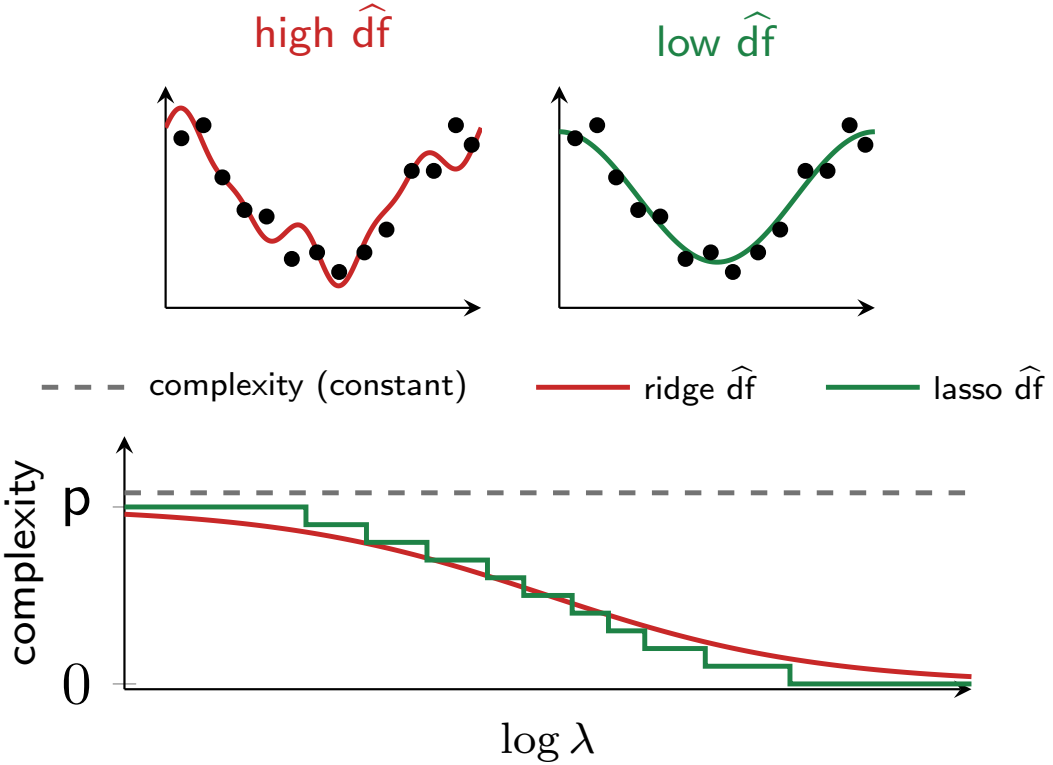


Caveat: SURE requires known or well-estimated σ^2 and a tractable df*; CV remains more general.*

Model Comparison with $\widehat{\text{df}}$

Unlike VC/Rademacher/Pollard complexities, which measure the worst-case property of the **class** $\mathcal{F}$, $\widehat{\text{df}}(\hat{f})$ is a property of the **specific fit** $\hat{f}$ produced.

- **Regularization is invisible to class measures.** OLS and heavy ridge share the same class complexity; $\widehat{\text{df}}(\lambda) = \text{tr}(H_\lambda)$ goes from p to 0 to tell them apart.
- **Common scale across families.** Ridge, lasso, splines, local regression, etc. are all from different classes; $\widehat{\text{df}}$ puts them on one *effective-parameter* axis.
- **Interpretability.** C_p, AIC, GCV, SURE all trade off fit against complexity via $\widehat{\text{df}}$, e.g. $\text{C}_\text{p} = \text{RSS} + 2\sigma^2\,\widehat{\text{df}}$.



Example: penalized regression

Calculating or estimating the degrees of freedom of estimators has been an active area of research (Ye, 1998; Zou et al., 2007; Hastie et al., 2009; Mazumder et al., 2011; Mazumder and Weng, 2020; Nobel et al., 2023).

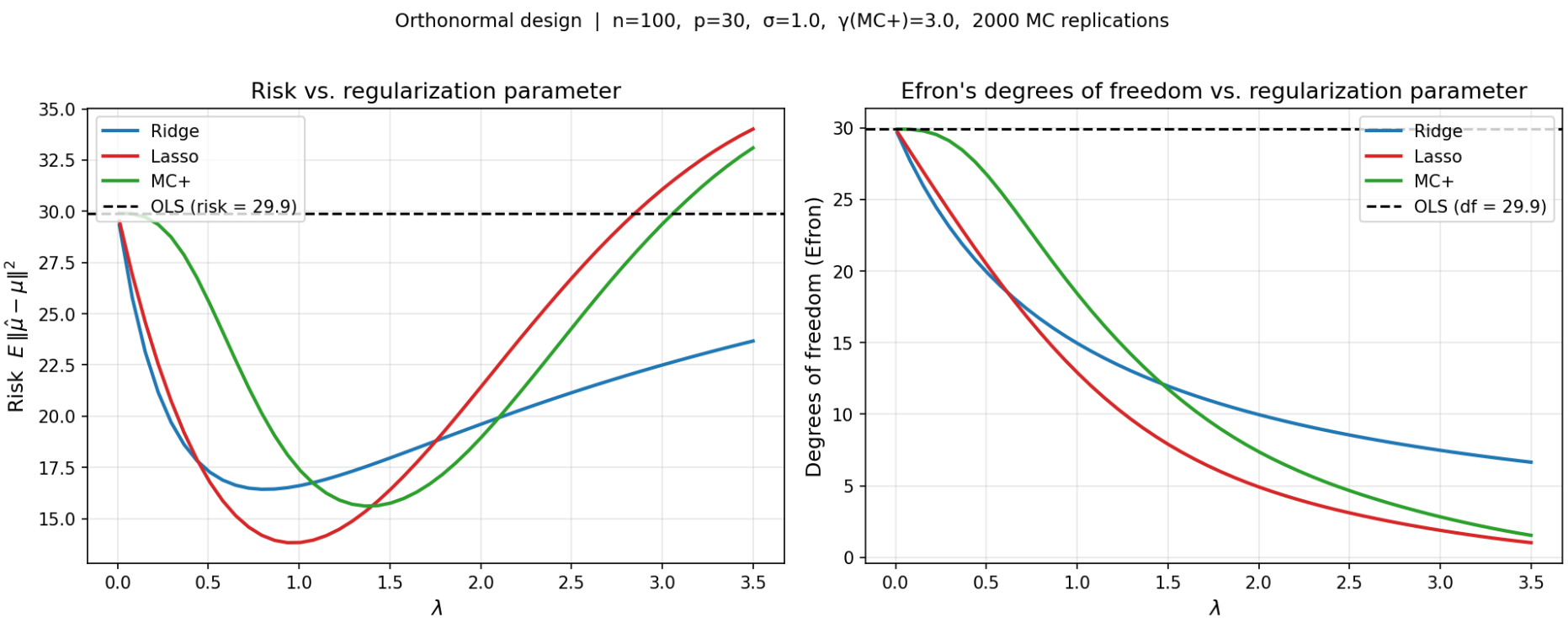


Figure: Risk curves for solutions of $\operatorname{argmin}_\beta \ \frac{1}{2}\|y - \beta x\|_2^2 + \lambda P(\beta)$, over differing $P(\beta)$.

Despite the problem being well-studied for regression, this is not the case for matrix approximation

The Matrix Approximation Problem

Consider $m \times n$ matrices signal X and noise $\mathcal{E}$, where we assume $\text{rank}(X) = r \ll k = \min\{n, m\}$,

$$Y = X + \mathcal{E}, \quad \text{iid. } \mathcal{E}_{ij} \text{ with } \mathbb{E}\mathcal{E}_{ij} = 0. \tag{1}$$

The goal is to recover X from the noisy observations Y.

This framework has been successfully used to tackle problems in

- Dimensionality Reduction & Visualization
- Recommendation Systems (Collaborative Filtering)
- Image Processing & Compression
- Latent Semantic Analysis (LSA)

Hard and Soft Singular Value Thresholding

Two of the most common ways of recovering the signal in (1) are Singular Value Hard-Thresholding (HSVT) and Soft-Thresholding (SVT).

$$(\mathsf{HSVT}): \quad \hat{f}^{\mathsf{HSVT}}(Y) = \underset{X}{\operatorname{argmin}} \frac{1}{2}\|Y - X\|_F^2 + \lambda \mathsf{rank}(X) = U_{(r)} D_{(r)} V_{(r)}^\top,$$

$$(\mathsf{SVT}): \quad \hat{f}^{\mathsf{SVT}}(Y) = \underset{X}{\operatorname{argmin}} \frac{1}{2}\|Y - X\|_F^2 + \lambda \|X\|_* = U[D - \lambda]_+ V^\top,$$

both expressed in terms of a Singular Value Decomposition of $Y = UDV^\top$, $D = \mathsf{diag}(d_1, \ldots, d_k)$.

> *When* $\mathcal{E}_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ *and* $\hat{f}$ *is weakly differentiable, it is possible to obtain* $\nabla \cdot \hat{f}$ *using Stein's lemma to estimate* $\mathsf{df}(\hat{f})$.

Interpretation for SVT

Theorem (*IV.3 in Candès et al. 2013*)

For a spectral estimator $\hat{f}$ of a simple matrix Y, its divergence is given by

$$\operatorname{div}(\hat{f}(Y)) = \underbrace{\sum_{i=1}^{k} \hat{f}_i'(d_i)}_{\text{(A)}} + \underbrace{|m-n| \sum_{i=1}^{k} \frac{\hat{f}_i(d_i)}{d_i}}_{\text{(B)}} + \underbrace{2 \sum_{i \neq j}^{k} \frac{d_i \hat{f}_i(d_i)}{d_i^2 - d_j^2}}_{\text{(C)}}.$$

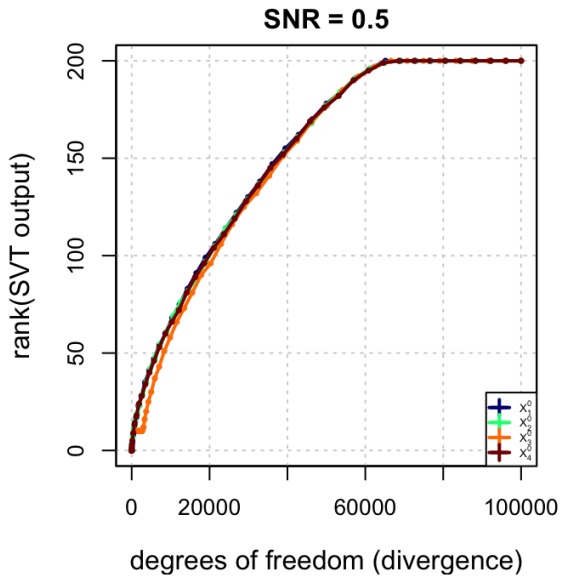
For SVT, the terms are

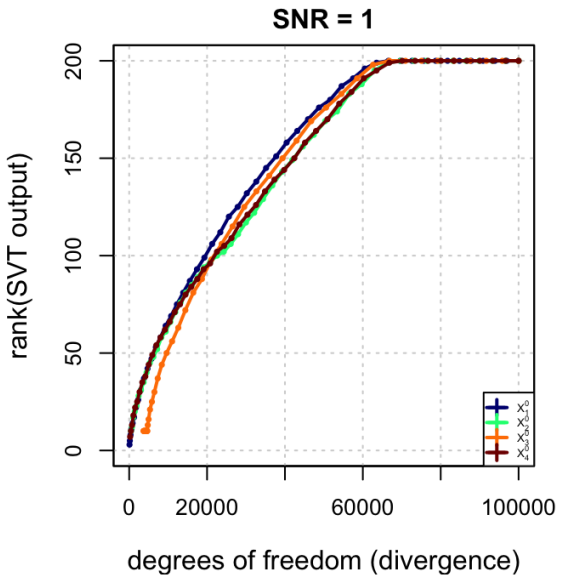
$$\text{(A)} = \sum_{i=1}^{k} \mathbb{I}(d_i > \lambda), \qquad \text{(B)} = |m-n| \sum_{i=1}^{k} \left(1 - \frac{\lambda}{d_i}\right)_+,$$

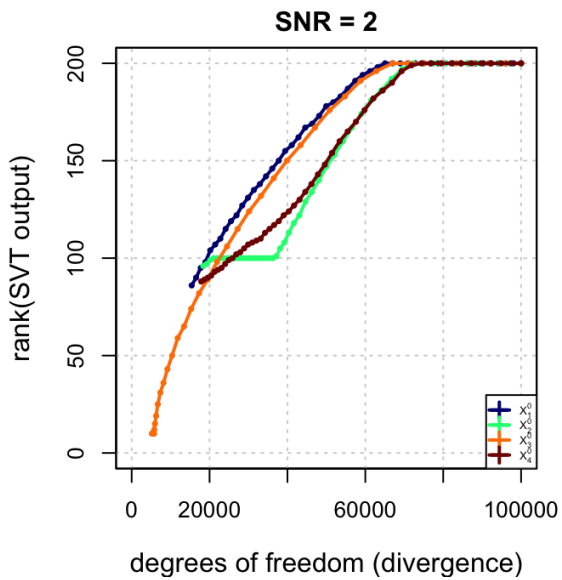
$$\text{(C)} = 2 \sum_{i \neq j}^{k} \frac{d_i (d_i - \lambda)_+}{d_i^2 - d_j^2}.$$

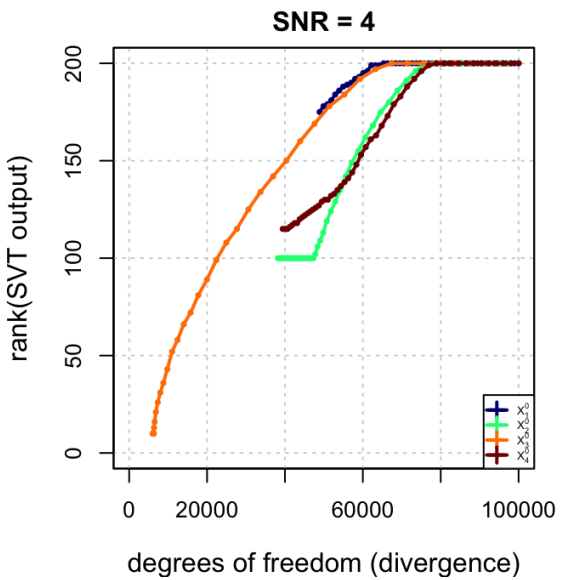
Proof idea. Leibniz rule for the matrix Jacobian on the SVD decomposition and then manipulate to decouple into univariate functions of the singular values. □

Effective Rank: Simulation for dfs of SVT









Regularized vector of singular values and Generalized Lasso

Many applications require regularizers of the entire vector of singular values.

$$\hat{f}_\lambda(Y) = \underset{X \in \mathbb{R}^{m\times n}}{\operatorname{argmin}} \frac{1}{2}\|Y - X\|_F^2 + \lambda Q(X),$$

to impose some structural sparsity on $\boldsymbol{d}$. Examples include:

- Multi-spike models
- Stochastic Block models
- Adjacency/Laplacian Matrices
- Trend filtering

Theorem (S. and Tuzhilina, 2026+))

When Q is a unitarily invariant function (only depends on the singular values of X), and $\hat{f}_\lambda$ is almost differentiable, then:

$$\operatorname{div}(\hat{f}(Y)) = \operatorname{tr}(\mathbf{J}_{\hat{f}}(\boldsymbol{d})) + |m-n| \sum_{i=1}^{k} \frac{\hat{f}_i(\boldsymbol{d})}{d_i} + 2\sum_{i\neq j} \frac{d_i \hat{f}_i(\boldsymbol{d})}{d_i^2 - d_j^2}. \qquad (2)$$

Example: Fused Lasso

The following is a Monte-Carlo simulation for a fused lasso-type penalty on the singular values,

$$\hat{f}(Y) = \underset{X}{\operatorname{argmin}} \ \frac{1}{2}\|Y - X\|_F^2 + \lambda \sum_{i=1}^{k-1} |d_i(X) - d_{i+1}(X)| \,.$$

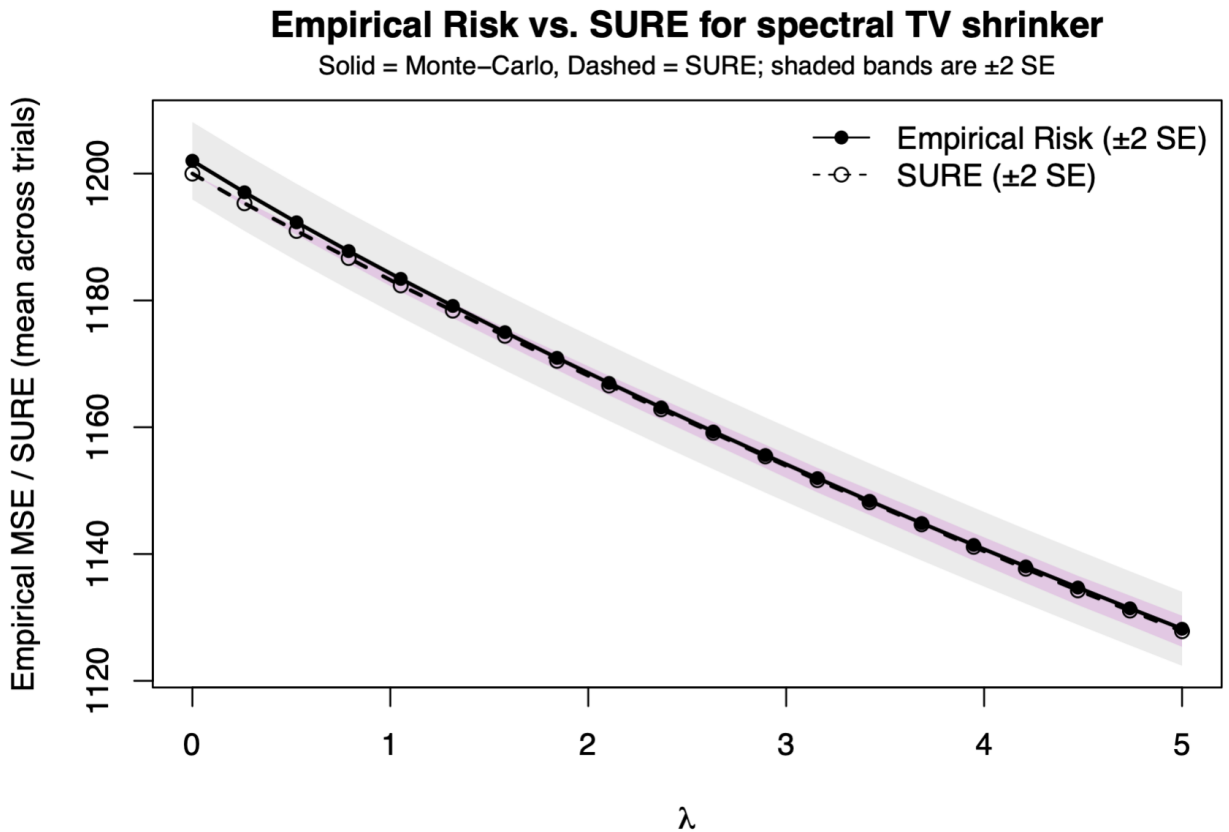


Figure: Comparison between SURE and MC for the TV shrinkage estimator in rank 5 random matrices of size 40×30.

Search Degrees of Freedom I

When $\hat{f}$ is not weakly differentiable, we cannot estimate the degrees of freedom using the divergence. An important example of this is Best Subset Selection; finding the best k variables on which to run a regression.

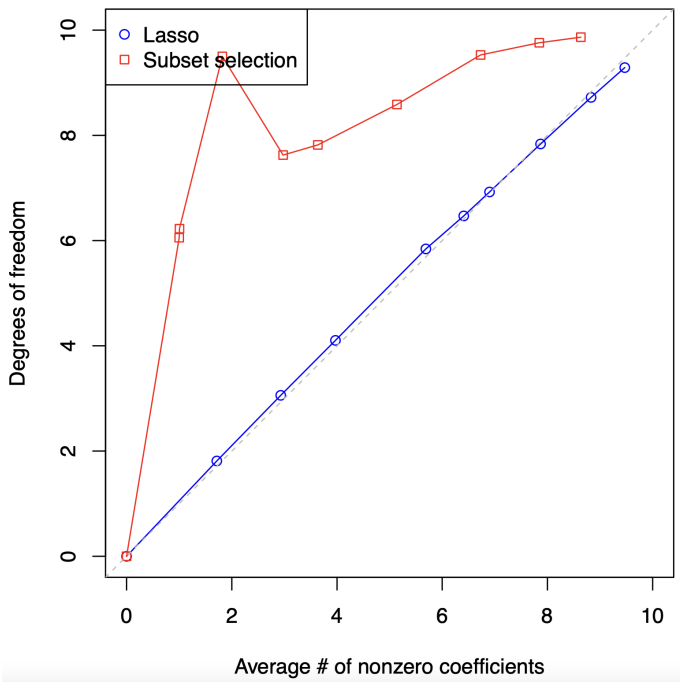


Figure: Simulated regression example with $n = 20, p = 10$ and fixed coefficients from Tibshirani (2015).

$$\hat{\beta}^{\text{subset}} \in \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_0,$$

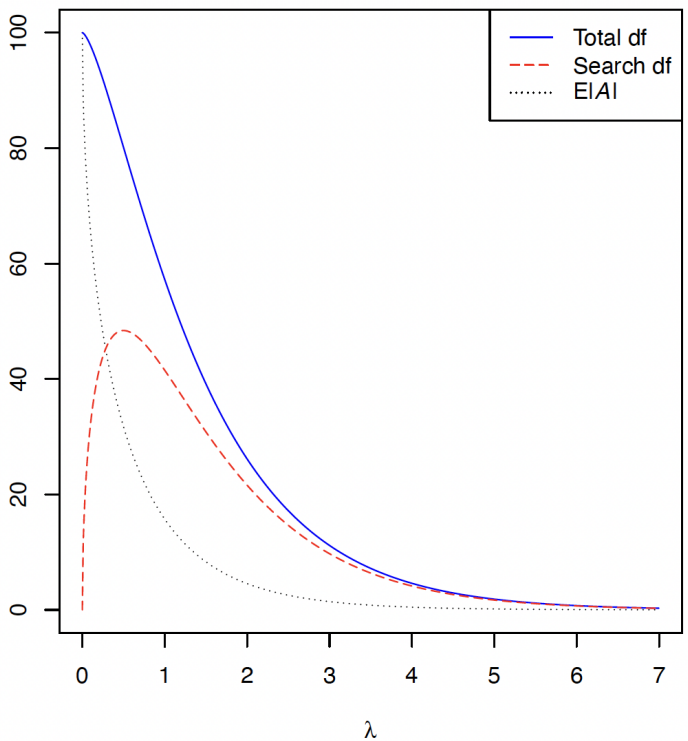
$$\text{df}(\hat{\beta}^{\text{subset}}) = \mathbb{E}|\mathcal{A}| + \text{sdf}(\hat{\beta}^{\text{subset}})$$

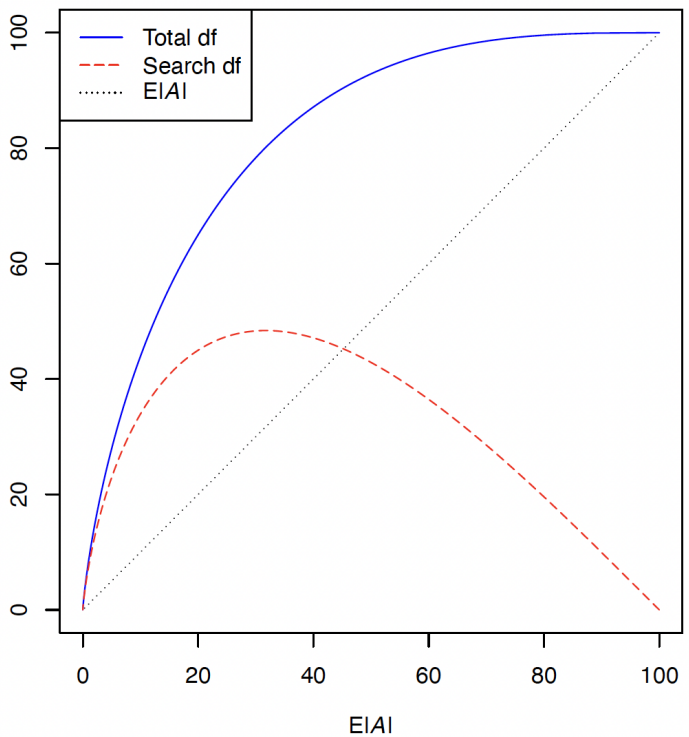
$$\text{df}(\hat{\beta}^{\text{lasso}}) = \mathbb{E}|\mathcal{A}|$$

Search Degrees of Freedom II

In the regression problem, the fitting cost of choosing the best k out of p coefficients on which to run the regression for orthogonal design X of the fit $\hat{\mu} = X\hat{\beta}^{\text{subset}}$ is (Tibshirani, 2015)

$$\mathsf{df}(\hat{\mu}) = \frac{\sqrt{2\lambda}}{\sigma} \sum_{i=1}^{p} \left[\phi\left(\frac{\sqrt{2\lambda} - (X^\top \mu)_i}{\sigma} \right) + \phi\left(\frac{\sqrt{2\lambda} + (X^\top \mu)_i}{\sigma} \right) \right].$$



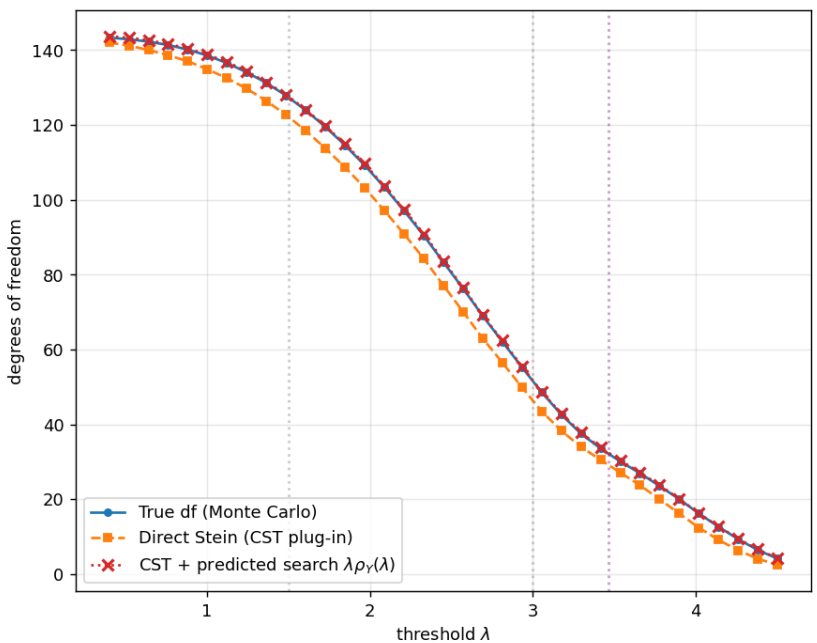


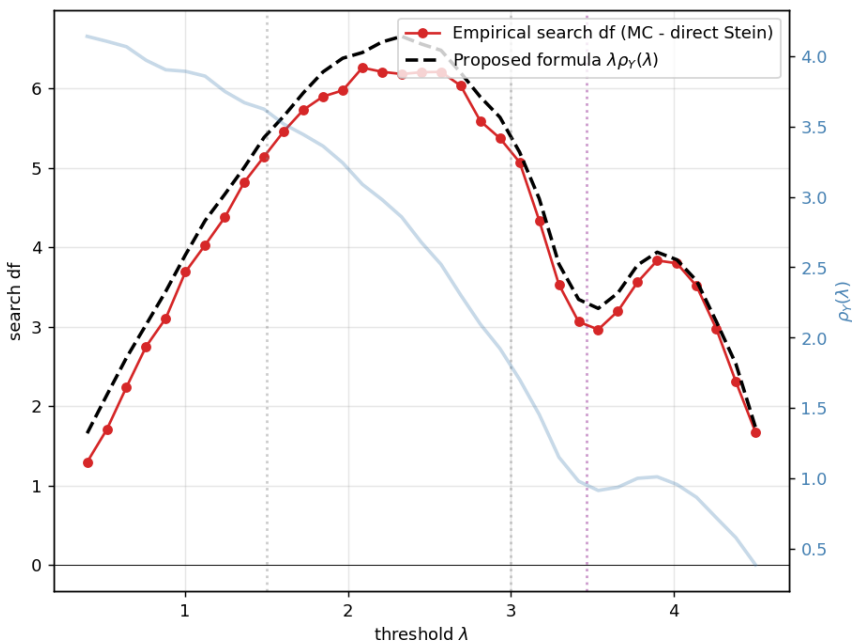
Search Degrees of Freedom III

Conjecture (S. and Tuzhilina (2026+))

Let $\rho_Y(\lambda)$ denote the spectral density of $Y^\top Y$. The search degrees of freedom for the HSVT estimator have expectation

$$\mathsf{sdf}(\hat{f}(Y)) = \lambda \rho_Y(\lambda) + \mathcal{O}(\lambda^{-1}), \quad \rho_Y(\lambda) = \sum_{i=1}^{k} f_{d_i(Y)}(\lambda).$$





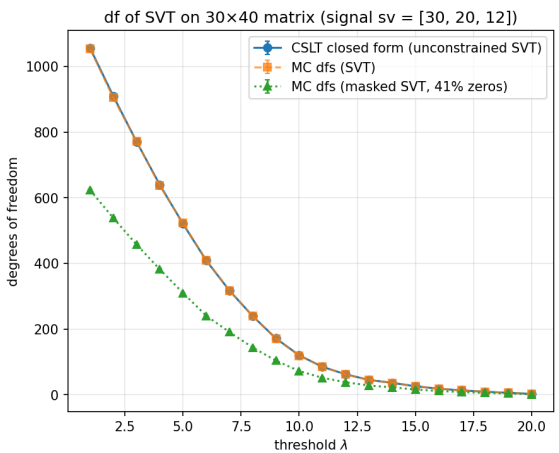
Sparsity Patterns

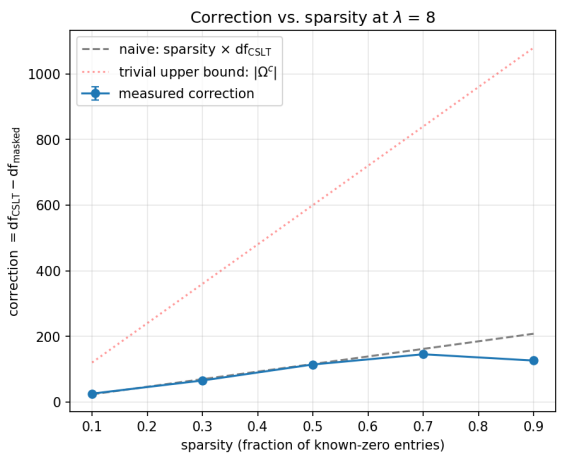
Let $\Omega \subset [m] \times [n]$ index the unknown entries and Ω^c the zero entries. Define $(P_\Omega A)_{ij} = A_{ij}\mathbf{1}_{(i,j)\in\Omega}$. The "plug-in" estimator $\hat{Y} = P_\Omega \circ \Phi^{\text{spec}}(\boldsymbol{Y})$ has a divergence,

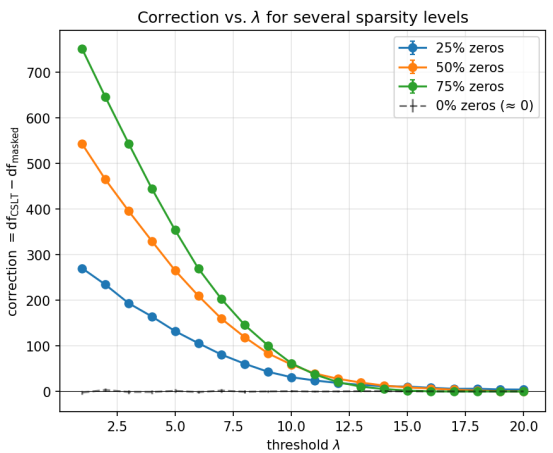
$$\text{div}\hat{Y} = \sum_{(i,j)\in\Omega} \partial_{Y_{ij}} \Phi^{\text{spec}}(\boldsymbol{Y})_{ij} = \underbrace{\text{div}\Phi^{\text{spec}}(\boldsymbol{Y})}_{\text{CSLT closed form}} - \underbrace{\sum_{(i,j)\in\Omega^c} \partial_{Y_{ij}} \Phi^{\text{spec}}(\boldsymbol{Y})_{ij}}_{\text{correction term}}.$$

Conjecture (S. and Tuzhilina (2026+))

A characterization of the effect of each entry on the final effective rank can be given by matrix perturbation theory.







Summary and future work

Summary:

- Degrees of freedom
- df also offer a way to analytically and unbiasedly estimate the risk of a sufficiently regular procedure.
- Even when the estimators are not regular, we can still find ways to bridge the gap (e.g. sdf).
- We can still derive more nice, closed form expressions for df of estimators $\hat{f}$, especially leveraging the known results from regression.
- The assumption of normality of errors is prevalent, and despite some extensions, these methods are still not robust to them.

Future work:

- Analytically derive df for other matrix approximation regularizers.
- Extend df to more complex models, or find suitable approximations.
- Improve our understanding of model fit!

References I

S. Bates, T. Hastie, and R. Tibshirani. Cross-validation: What does it estimate and how well does it do it? *Journal of the American Statistical Association*, 119(546):1434–1445, 2023. doi: 10.1080/01621459.2023.2197686.

E. J. Candès, C. A. Sing-Long, and J. D. Trzasko. Unbiased risk estimates for singular value thresholding and spectral estimators. *IEEE Transactions on Signal Processing*, 61(19): 4643–4657, 2013. doi: 10.1109/TSP.2013.2270464.

E. J. Candès, C. A. Sing-Long, and J. D. Trzasko. Unbiased Risk Estimates for Singular Value Thresholding and Spectral Estimators. *IEEE Transactions on Signal Processing*, 61(19): 4643–4657, Oct. 2013. ISSN 1053-587X, 1941-0476. doi: 10.1109/TSP.2013.2270464. URL `http://ieeexplore.ieee.org/document/6545395/`.

B. Efron. How biased is the apparent error rate of a prediction rule? *Journal of the American Statistical Association*, 81(394):461–470, 1986. doi: 10.1080/01621459.1986.10478291.

B. Efron. The estimation of prediction error: Covariance penalties and cross-validation. *Journal of the American Statistical Association*, 99(467):619–632, 2004. doi: 10.1198/016214504000000692.

Y. C. Eldar. Generalized sure for exponential families: Applications to regularization. *IEEE Transactions on Signal Processing*, 57(2):471–481, 2008. doi: 10.1109/TSP.2008.2008212.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer New York, New York, NY, second edition, 2009. doi: 10.1007/978-0-387-84858-7.

References II

R. Mazumder and H. Weng. Computing the degrees of freedom of rank-regularized estimators and cousins. *Electronic Journal of Statistics*, 14(1):1348 – 1385, 2020. doi: 10.1214/20-EJS1681. URL `https://doi.org/10.1214/20-EJS1681`.

R. Mazumder, J. Friedman, and T. Hastie. Sparsenet: Coordinate descent with nonconvex penalties. *Journal of the American Statistical Association*, 106(495):1125–1138, 2011.

P. Nobel, E. Candès, and S. Boyd. Tractable evaluation of stein's unbiased risk estimate with convex regularizers. *IEEE Transactions on Signal Processing*, 71:4330–4341, 2023. doi: 10.1109/tsp.2023.3323046.

C. M. Stein. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135–1151, 1981. doi: 10.1214/aos/1176345632.

R. J. Tibshirani. Degrees of freedom and model search. *Statist. Sinica*, 25:1265–1296, 2015.

R. J. Tibshirani and L. Wasserman. Stein's unbiased risk estimate. Lecture notes for Statistical Machine Learning (10-702/36-702), Carnegie Mellon University, Spring 2015. URL `https://www.stat.cmu.edu/~larry/=sml/stein.pdf`.

J. Ye. On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association*, 93(441):120–131, 1998. doi: 10.1080/01621459.1998.10474094.

H. Zou, T. Hastie, and R. Tibshirani. On the "degrees of freedom" of the lasso. *The Annals of Statistics*, 35(5):2173–2192, 2007. doi: 10.1214/009053607000000127.

Simulations

Synthetic 200×500 matrices reproducing Candès et al. (2013):

- X_1^0: full column rank with entries $X_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$.
- X_2^0: same as X_2^0 truncating down to the first 100 singular values.
- X_3^0: same as X_2^0 truncating down to the first 10 singular values.
- X_4^0: s. values $d_i = \sqrt{200}/\left(1 + e^{(i-100)/20}\right)$, for $i \in [200]$, and random s.vectors.

We compare the use of Monte Carlo estimate of the risk

$$\widehat{\text{Risk}}_{\text{MC}}(f_\lambda) = \frac{1}{50}\sum_{j=1}^{50} \|X_i^0 - \text{SVT}_\lambda(Y_j^{(i)})\|_F^2$$

and the SURE estimator

$$\text{SURE}(\text{SVT}_\lambda(Y^{(i)})) = -mn\tau^2 + \sum_{i=1}^{k} \min\{\lambda^2, d_i^2\} + 2\tau^2 \text{div}\left(\text{SVT}_\lambda(Y^{(i)})\right).$$

Simulations II

Fixing SNR $= \|X_i^0\|_F^2/\sqrt{nm}\tau = 1/\sqrt{nm}\tau \in \{0.5, 1, 2, 4\}$