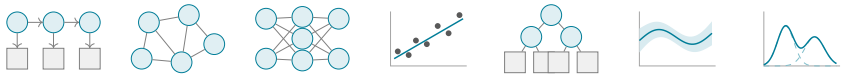
Degrees of freedom and the model selection problem

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

Joint Mathematics-Statistics Colloquium
University of Toronto
November 2026

A new problem. Which model?



When choosing a model, statisticians do one of the following things:

- Look at what models are usually chosen for that task.
- Choose a model that "makes sense" given the nature of the problem.
- Try to predict which model will have the *best performance*.

Setting: Signal + Noise Model

We observe n responses generated from a fixed mean:

$$\mathbf{y} = \boldsymbol{\mu}(x) + \boldsymbol{\varepsilon}, \qquad \mathbb{E}[\boldsymbol{\varepsilon}] = \mathbf{0}, \quad \mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}_n.$$

- $\mu = (\mu_1, \ldots, \mu_n)^\top$ is the *true mean* at the design points, $\mu_i = \mathbb{E}[y_i \mid x]$
- ε is mean-zero, iid. noise with common variance σ^2.
- A fitting procedure turns the data into predictions $\hat{y} = \hat{f}(y)$ at the same n points.

Predicting how well we predict

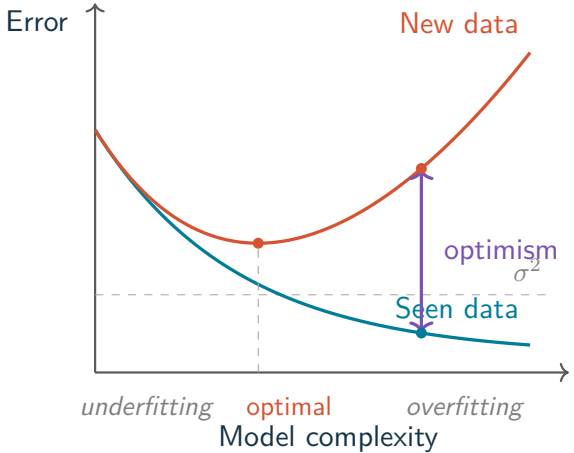
Suppose we have some set of statistical predictors to choose from $\mathcal{P} = \left\{\hat{f}_k\right\}_k$

$$y \xrightarrow{\text{fit}} \hat{f}_k(y), \qquad R_k := \mathbb{E}\left\|\hat{f}_k(Y) - \mu\right\|^2$$

$$y \xrightarrow{\text{estimate risk}} \widehat{R}_k(y) \approx R_k$$

$$\boxed{\hat{f} \in \underset{\hat{f}_k \in \mathcal{P}}{\arg\min}\ \hat{R}_k(y)} \qquad \Longrightarrow \qquad \hat{f}_{\widehat{k}}(y)$$

Prediction error and optimism



SEEN RESPONSES

$$L_{\text{train}} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}_y(x_i))^2$$

NEW RESPONSES

$$L_{\text{new}} = \frac{1}{n}\sum_{i=1}^{n}(y_i^{\text{new}} - \hat{f}_y(x_i))^2$$

$$\text{Opt}(\hat{f}) = \mathbb{E}[L_{\text{new}} \mid X] - \mathbb{E}[L_{\text{train}} \mid X]$$

Risk estimation and Stein's lemma

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

$$\mathrm{R}(\hat{f}) = \mathbb{E}\|\mu - \hat{f}\|_2^2 = \mathbb{E}\|y - \hat{f}(x)\|_2^2 - n\sigma^2 + 2\sigma^2 \mathsf{df}(\hat{f}),$$
$$\mathsf{df}(\hat{f}) := \frac{1}{\sigma^2} \sum_{i=1}^{n} \mathrm{Cov}\Big(\hat{f}(x_i), y_i\Big).$$

Stein's lemma

Let $y \sim \mathcal{N}(\mu, \sigma^2 I_n)$. If $\hat{f} : \mathbb{R}^n \to \mathbb{R}^n$ is **weakly differentiable**, with square-integrable fitted values and integrable diagonal derivatives, then

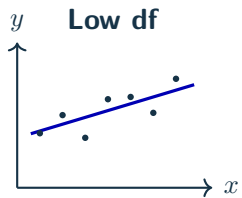
$$\boxed{\mathsf{df}(\hat{f}) = \mathbb{E}\Big[\mathrm{tr}\left(D\hat{f}(y)\right)\Big].}$$

What happens when $\hat{f}$ is not weakly differentiable?

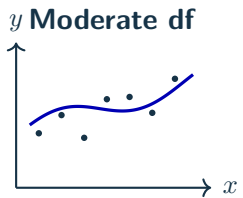
Degrees of Freedom

$$\mathrm{df}(\hat{f}) \;=\; \frac{1}{\sigma^2}\sum_{i=1}^{n}\mathrm{Cov}\left(\hat{f}(y)_i, y_i\right).$$

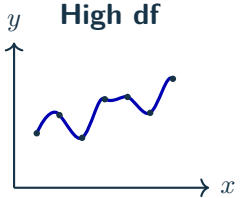
- Compared to AIC or BIC, this definition is likelihood free.
- If $\hat{f}$ ignores the data, every covariance is $0 \Rightarrow \mathrm{df} = 0$.
- If $\hat{f}$ interpolates the data ($\hat{f}(y_i) = y_i$), then $\mathrm{Cov}(\hat{f}(y_i), y_i) = \sigma^2 \Rightarrow \mathrm{df} = n$.



df = 2 (linear fit)



df ≈ 4 (smoother fit)



df = n = 7 (interpolation)

Example: Linear Regression

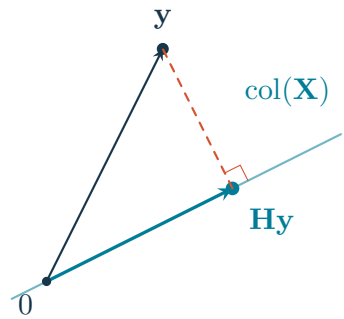
In OLS, the fit is linear in the response:

$$\hat{f}(\mathbf{y}) = \mathbf{H}\mathbf{y}, \qquad \mathbf{H} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top.$$

With fixed predictors, $\mathbf{H}$ does not change with $\mathbf{Y}$. Therefore,

$$\mathrm{Cov}(\hat{f}(\mathbf{Y})_i, Y_i) = \sigma^2 H_{ii},$$

$$\mathsf{df}(\hat{f}) = \sum_{i=1}^{n} H_{ii} = \mathrm{tr}(\mathbf{H}).$$



$\mathbf{H}$ is the orthogonal projection onto $\mathrm{col}(\mathbf{X})$, so

$$\boxed{\mathsf{df}(\hat{f}) = \mathrm{tr}(\mathbf{H}) = \mathrm{rank}(\mathbf{X}) = p}$$

fitted parameters.

Assumptions: $\mathrm{rank}(\mathbf{X}) = p$ and $\mathrm{Cov}(\mathbf{Y}) = \sigma^2 I_n$.

Example: penalized regression

Estimating dfs 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).

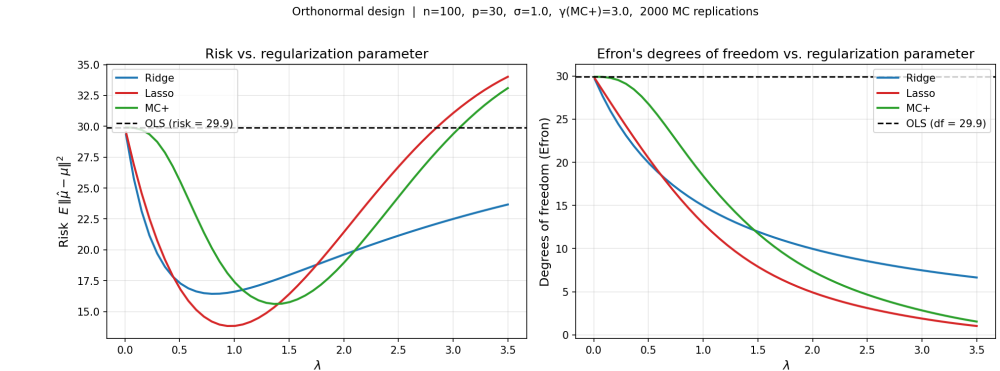
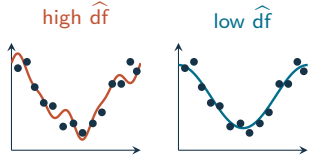


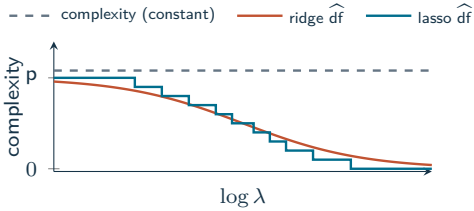
Figure: Risk curves for solutions of $\arg\min_\beta \ \frac{1}{2}\|y - x^\top \beta\|_2^2 + \lambda P(\beta)$, over differing $P(\beta)$.

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

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

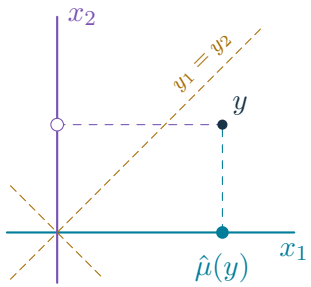
- **Regularization is invisible to class measures.** OLS and heavy ridge share the same class complexity; $\widehat{\mathrm{df}}(\lambda) = \mathrm{tr}(H_\lambda)$ goes from p to 0 to tell them apart.
- **Common scale across families.** Ridge, lasso, splines, local regression, etc. are all from different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





Example: Best subset selection

We now have two orthogonal predictors x_1, x_2 and want to select one by least squares.



$$(y_1, y_2) \sim \mathcal{N}(0, I_2), \qquad K = \{x_2 = 0\} \cup \{x_2 = 0\}\,.$$

$$\hat{f}(y) \in \arg\min_{z \in K} \|y - z\|_2^2.$$

$$\hat{f}(y) = \begin{cases} (y_1, 0), & |y_1| \geq |y_2|, \\ (0, y_2), & |y_2| > |y_1|. \end{cases}$$

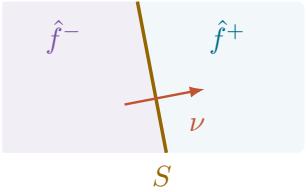
In either open selection region,

$$\operatorname{tr}\Big(D\hat{f}(y)\Big) = 1.$$

$$\mathsf{df}(\hat{f}) = \mathbb{E}\left[\max\{y_1^2, y_2^2\}\right] = 1 + \frac{2}{\pi} \approx 1.64.$$

The fit jumps across $|y_1| = |y_2|$.

Search degrees of freedom



Model selection regions in data space

S: interfaces where the selected model changes.

$[\hat{f}] = \hat{f}^+ - \hat{f}^-, \qquad \nu$: unit normal from $-$ to $+$.

Integration by parts in each region leaves a boundary term whenever the fitted values jump across S.

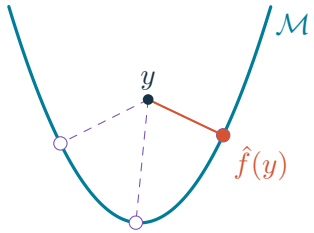
$$\mathsf{df}(\hat{f}) = \underbrace{\mathbb{E}\Big[\mathrm{tr}(D\hat{f}(Y))\Big]}_{\text{fitting the selected model}} + \underbrace{\int_S \langle [\hat{f}](y), \nu(y)\rangle \, p_{\mu,\sigma}(y)\, d\mathcal{H}^{n-1}(y)}_{\text{search degrees of freedom } \mathsf{sdf}(\hat{f})}.$$

In our two-axis example: $\mathsf{df}(\hat{f}) = 1 + \mathsf{sdf}(\hat{f}) = 1 + \dfrac{2}{\pi}$.

$p_{\mu,\sigma}$: Gaussian density. $\mathcal{H}^{n-1}$: surface measure. Each interface counted once.

Degrees of freedom and algebraic geometry

$$\mathcal{M} = \{(u, v) \in \mathbb{R}^2 : v - u^2 = 0\} = \{(t, t^2) : t \in \mathbb{R}\}.$$



Three real critical points shown

$$\operatorname{tr}\Big(D\hat{f}(a,b)\Big) = \frac{1 + 4\hat{t}^2}{6\hat{t}^2 + 1 - 2b}.$$

For $y = (a, b)$, fit $\hat{f}(y) = (\hat{t}, \hat{t}^2)$:

$$\hat{t} \in \operatorname*{arg\,min}_{t \in \mathbb{R}} \big[(t - a)^2 + (t^2 - b)^2\big].$$

The residual is normal to the tangent:

$$\Big\langle (t, t^2) - (a, b),\ (1, 2t) \Big\rangle = 0.$$

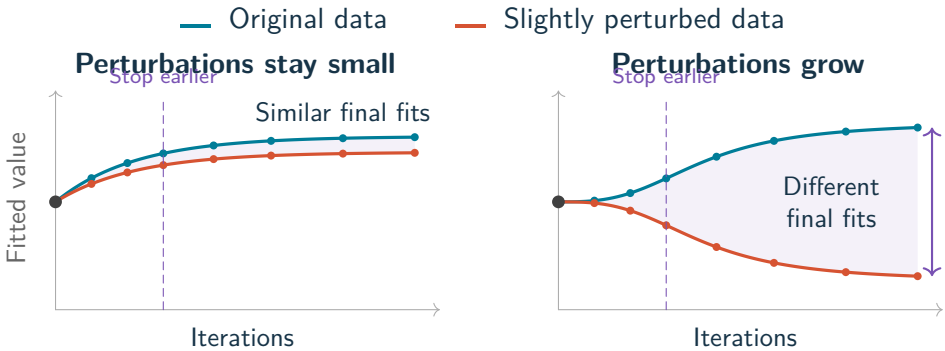
The critical equation is cubic:

$$\boxed{2t^3 + (1 - 2b)t - a = 0}.$$

$$\dim(\mathcal{M}) = 1, \qquad \mathrm{EDdeg}(\mathcal{M}) = 3.$$

Euclidean distance degree: the number of complex critical points for generic data.

Degrees of freedom and iterative dynamics



Degrees of freedom: how much the fit follows the data.

The iteration rule and stopping time are part of the estimator.

Main Takeaways

- Degrees of freedom measure adaptation to the data, extending classical parameter counting.
- Degrees of freedom can be used as a measure of model complexity
- df for unbiased risk estimation.
- 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.

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.

J. Draisma, E. Horobeţ, G. Ottaviani, B. Sturmfels, and R. R. Thomas. The Euclidean distance degree of an algebraic variety. *Foundations of Computational Mathematics*, 16(1):99–149, 2016. doi: 10.1007/s10208-014-9240-x.

L. C. Evans and R. F. Gariepy. *Measure Theory and Fine Properties of Functions*. Textbooks in Mathematics. CRC Press, Boca Raton, FL, revised edition, 2015. ISBN 9781482242386.

N. R. Hansen. On Stein's unbiased risk estimate for reduced rank estimators. *Statistics & Probability Letters*, 135:76–82, 2018. doi: 10.1016/j.spl.2017.11.006.

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.

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.

References II

F. R. Mikkelsen and N. R. Hansen. Degrees of freedom for piecewise Lipschitz estimators. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 54(2):819–841, 2018. doi: 10.1214/17-AIHP822.

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.

P. Patil, J.-H. Du, and R. J. Tibshirani. Revisiting optimism and model complexity in the wake of overparameterized machine learning. *arXiv*, 2024. doi: 10.48550/arxiv.2410.01259.

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

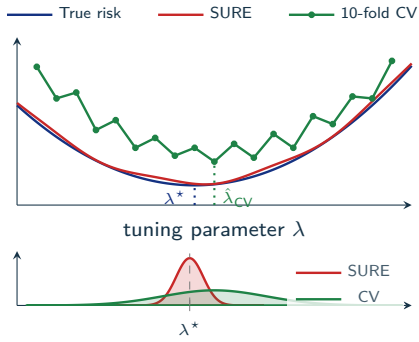
J. Ye. On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association*, 93(441):120–131, 1998. doi: 10.1080/01621459.1998.10474094.

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.

How 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).



selected $\hat{\lambda}$ across replicates, for $\hat{f} = \hat{f}_{\hat{\lambda}}$.

Overparametrized Learning

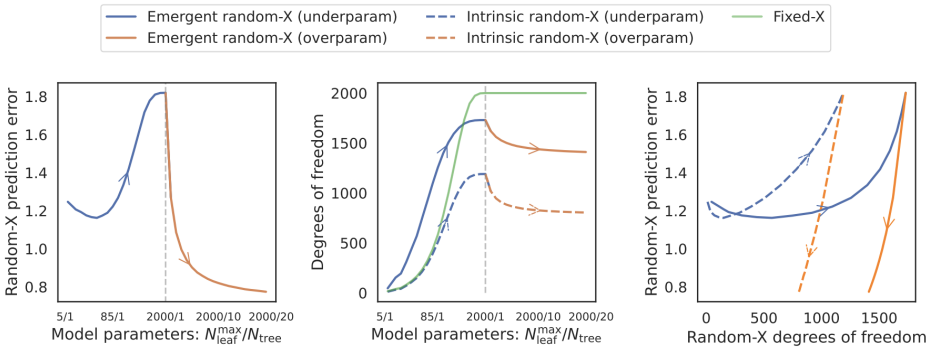


Figure: Figure from Patil et al. (2024) showing the prediction error, degrees of freedom of random forest predictors as the number of trees N_{tree} and the maximum number of leaves $N_{\text{leaf}}^{\text{max}}$ is changed.

Example: Neural Networks

