Facets of regularization in high-dimensional learning:

Cross-validation, risk monotonization, and model complexity

Pratik Patil

Carnegie Mellon University

November 2022

Committee:

Ryan Tibshirani (Chair)
Alessandro Rinaldo
Arun Kumar Kuchibhotla
Yuting Wei (University of Pennsylvania)
Arian Maleki (Columbia University)

Outline

Overparametrization in machine learning

Modern machine learning models typically fit a huge number of parameters. Such overparameterization seems to be useful for[1]:

- Representation: allows rich, expressive models for diverse real data
- Optimization: simple, local optimization methods often find near-optimal solutions to empirical risk minimization problem
- Generalization: despite overfitting, models generalize well in practice

This talk is about generalization aspect in overparameterized learning.

[1] Credits to Ryan for this nice partition of distinct benefits of overparameterization.

Overparametrization in machine learning

Modern machine learning models typically fit a huge number of parameters. Such overparameterization seems to be useful for[1]:

- Representation: allows rich, expressive models for diverse real data
- Optimization: simple, local optimization methods often find near-optimal solutions to empirical risk minimization problem
- Generalization: despite overfitting, models generalize well in practice

This talk is about generalization aspect in overparameterized learning.

[1] Credits to Ryan for this nice partition of distinct benefits of overparameterization.

Overparametrization in machine learning

Modern machine learning models typically fit a huge number of parameters. Such overparameterization seems to be useful for[1]:

- Representation: allows rich, expressive models for diverse real data
- Optimization: simple, local optimization methods often find near-optimal solutions to empirical risk minimization problem
- Generalization: despite overfitting, models generalize well in practice

This talk is about generalization aspect in overparameterized learning.

[1] Credits to Ryan for this nice partition of distinct benefits of overparameterization.

Overparametrization in machine learning

Modern machine learning models typically fit a huge number of parameters. Such overparameterization seems to be useful for[1]:

- Representation: allows rich, expressive models for diverse real data
- Optimization: simple, local optimization methods often find near-optimal solutions to empirical risk minimization problem
- Generalization: despite overfitting, models generalize well in practice

This talk is about generalization aspect in overparameterized learning.

[1] Credits to Ryan for this nice partition of distinct benefits of overparameterization.

Overparametrization in machine learning

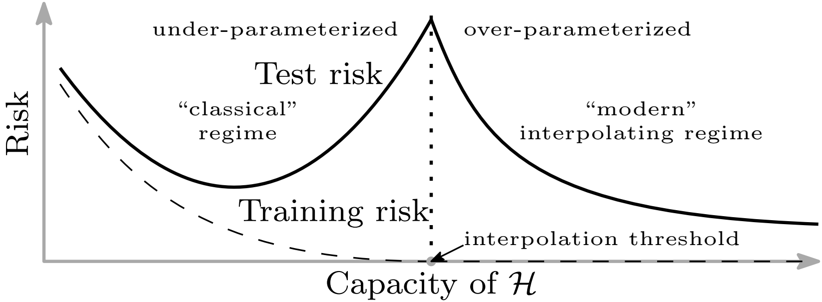
Modern machine learning models typically fit a huge number of parameters. Such overparameterization seems to be useful for[1]:

- Representation: allows rich, expressive models for diverse real data
- Optimization: simple, local optimization methods often find near-optimal solutions to empirical risk minimization problem
- Generalization: despite overfitting, models generalize well in practice

This talk is about generalization aspect in overparameterized learning.

[1] Credits to Ryan for this nice partition of distinct benefits of overparameterization.

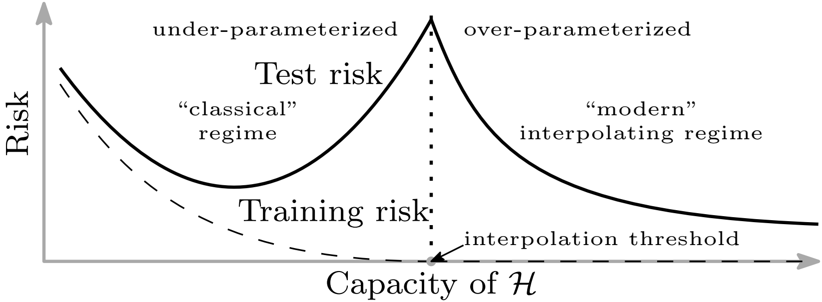
Peculiar generalization behavior: double descent



Belkin, Hsu, Ma, Mandal, 2018: "Reconciling modern machine learning practice and the bias variance tradeoff"

- The phenomenon is dubbed "double descent" in the risk curve.
- This trend holds for many model classes including linear regression, kernel regression, random forest, boosting, neural networks, etc.

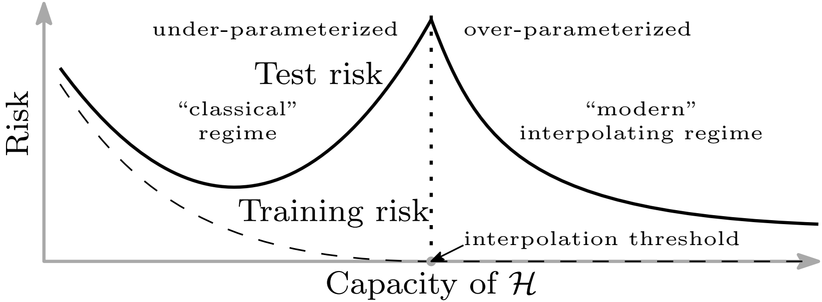
Peculiar generalization behavior: double descent



Belkin, Hsu, Ma, Mandal, 2018: "Reconciling modern machine learning practice and the bias variance tradeoff"

- The phenomenon is dubbed "double descent" in the risk curve.
- This trend holds for many model classes including linear regression, kernel regression, random forest, boosting, neural networks, etc.

Peculiar generalization behavior: double descent



Belkin, Hsu, Ma, Mandal, 2018: "Reconciling modern machine learning practice and the bias variance tradeoff"

- The phenomenon is dubbed "double descent" in the risk curve.
- This trend holds for many model classes including linear regression, kernel regression, random forest, boosting, neural networks, etc.

Recent theoretical developments

Understanding generalization of interpolators in simpler settings:

- Linear regression
  - Hastie, Montanari, Rosset, Tibshirani, 2019
  - Belkin, Hsu, Xu, 2019
  - Muthukumar, Vodrahalli, Sahai, 2019
  - Bartlett, Long, Lugosi, Tsigler, 2019
  - Mei, Montanari, 2019
- Kernel regression
  - Liang, Rakhlin, 2018
  - Liang, Rakhlin, Zhai, 2019
- Local methods
  - Belkin, Hsu, Mitra, 2018
  - Belkin, Rakhlin, Tsybakov, 2018
- and many more ...

Nice survey papers:

- Bartlett, Montanari, and Rakhlin, 2021: "Deep learning: a statistical viewpoint"
- Belkin, 2021: "Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation"

Recent theoretical developments

Understanding generalization of interpolators in simpler settings:

- Linear regression
  - Hastie, Montanari, Rosset, Tibshirani, 2019
  - Belkin, Hsu, Xu, 2019
  - Muthukumar, Vodrahalli, Sahai, 2019
  - Bartlett, Long, Lugosi, Tsigler, 2019
  - Mei, Montanari, 2019
- Kernel regression
  - Liang, Rakhlin, 2018
  - Liang, Rakhlin, Zhai, 2019
- Local methods
  - Belkin, Hsu, Mitra, 2018
  - Belkin, Rakhlin, Tsybakov, 2018
- and many more ...

Nice survey papers:

- Bartlett, Montanari, and Rakhlin, 2021: "Deep learning: a statistical viewpoint"
- Belkin, 2021: "Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation"

Recent theoretical developments

Understanding generalization of interpolators in simpler settings:

- Linear regression
  - Hastie, Montanari, Rosset, Tibshirani, 2019
  - Belkin, Hsu, Xu, 2019
  - Muthukumar, Vodrahalli, Sahai, 2019
  - Bartlett, Long, Lugosi, Tsigler, 2019
  - Mei, Montanari, 2019
- Kernel regression
  - Liang, Rakhlin, 2018
  - Liang, Rakhlin, Zhai, 2019
- Local methods
  - Belkin, Hsu, Mitra, 2018
  - Belkin, Rakhlin, Tsybakov, 2018
- and many more ...

Nice survey papers:

- Bartlett, Montanari, and Rakhlin, 2021: "Deep learning: a statistical viewpoint"
- Belkin, 2021: "Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation"

Recent theoretical developments

Understanding generalization of interpolators in simpler settings:

- Linear regression
  - Hastie, Montanari, Rosset, Tibshirani, 2019
  - Belkin, Hsu, Xu, 2019
  - Muthukumar, Vodrahalli, Sahai, 2019
  - Bartlett, Long, Lugosi, Tsigler, 2019
  - Mei, Montanari, 2019
- Kernel regression
  - Liang, Rakhlin, 2018
  - Liang, Rakhlin, Zhai, 2019
- Local methods
  - Belkin, Hsu, Mitra, 2018
  - Belkin, Rakhlin, Tsybakov, 2018
- and many more ...

Nice survey papers:

- Bartlett, Montanari, and Rakhlin, 2021: "Deep learning: a statistical viewpoint"
- Belkin, 2021: "Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation"

Recent theoretical developments

Understanding generalization of interpolators in simpler settings:

- Linear regression
  - Hastie, Montanari, Rosset, Tibshirani, 2019
  - Belkin, Hsu, Xu, 2019
  - Muthukumar, Vodrahalli, Sahai, 2019
  - Bartlett, Long, Lugosi, Tsigler, 2019
  - Mei, Montanari, 2019
- Kernel regression
  - Liang, Rakhlin, 2018
  - Liang, Rakhlin, Zhai, 2019
- Local methods
  - Belkin, Hsu, Mitra, 2018
  - Belkin, Rakhlin, Tsybakov, 2018
- and many more ...

Nice survey papers:

- Bartlett, Montanari, and Rakhlin, 2021: "Deep learning: a statistical viewpoint"
- Belkin, 2021: "Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation"

Recent theoretical developments

Understanding generalization of interpolators in simpler settings:

- Linear regression
  - Hastie, Montanari, Rosset, Tibshirani, 2019
  - Belkin, Hsu, Xu, 2019
  - Muthukumar, Vodrahalli, Sahai, 2019
  - Bartlett, Long, Lugosi, Tsigler, 2019
  - Mei, Montanari, 2019
- Kernel regression
  - Liang, Rakhlin, 2018
  - Liang, Rakhlin, Zhai, 2019
- Local methods
  - Belkin, Hsu, Mitra, 2018
  - Belkin, Rakhlin, Tsybakov, 2018
- and many more ...

Nice survey papers:

- Bartlett, Montanari, and Rakhlin, 2021: "Deep learning: a statistical viewpoint"
- Belkin, 2021: "Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation"

Motivating questions

We study three operational aspects of overparameterized learning:
1) cross-validation, 2) risk monotonization, 3) model complexity.

Motivating questions:

1. Does cross-validation still "work" in the overparameterized regime, especially when optimal regularization and train error can be zero?

2. Is it possible to modify any given prediction procedure to mitigate double descent behavior and achieve a monotonic risk behavior?

3. Is there a better and more principled measure of model complexity in general for overparameterized models?

Short answers: YES.

Long answers: Rest of the talk.

Motivating questions

We study three operational aspects of overparameterized learning:
1) cross-validation, 2) risk monotonization, 3) model complexity.

Motivating questions:

1. Does cross-validation still "work" in the overparameterized regime, especially when optimal regularization and train error can be zero?

2. Is it possible to modify any given prediction procedure to mitigate double descent behavior and achieve a monotonic risk behavior?

3. Is there a better and more principled measure of model complexity in general for overparameterized models?

Short answers: YES.

Long answers: Rest of the talk.

Motivating questions

We study three operational aspects of overparameterized learning:
1) cross-validation, 2) risk monotonization, 3) model complexity.

Motivating questions:

1. Does cross-validation still "work" in the overparameterized regime, especially when optimal regularization and train error can be zero?

2. Is it possible to modify any given prediction procedure to mitigate double descent behavior and achieve a monotonic risk behavior?

3. Is there a better and more principled measure of model complexity in general for overparameterized models?

Short answers: YES.

Long answers: Rest of the talk.

Motivating questions

We study three operational aspects of overparameterized learning:
1) cross-validation, 2) risk monotonization, 3) model complexity.

Motivating questions:

1. Does cross-validation still "work" in the overparameterized regime, especially when optimal regularization and train error can be zero?

2. Is it possible to modify any given prediction procedure to mitigate double descent behavior and achieve a monotonic risk behavior?

3. Is there a better and more principled measure of model complexity in general for overparameterized models?

Short answers: YES.

Long answers: Rest of the talk.

Motivating questions

We study three operational aspects of overparameterized learning:
1) cross-validation, 2) risk monotonization, 3) model complexity.

Motivating questions:

1. Does cross-validation still "work" in the overparameterized regime, especially when optimal regularization and train error can be zero?

2. Is it possible to modify any given prediction procedure to mitigate double descent behavior and achieve a monotonic risk behavior?

3. Is there a better and more principled measure of model complexity in general for overparameterized models?

Short answers: YES.

Long answers: Rest of the talk.

Motivating questions

We study three operational aspects of overparameterized learning:
1) cross-validation, 2) risk monotonization, 3) model complexity.

Motivating questions:

1. Does cross-validation still "work" in the overparameterized regime, especially when optimal regularization and train error can be zero?

2. Is it possible to modify any given prediction procedure to mitigate double descent behavior and achieve a monotonic risk behavior?

3. Is there a better and more principled measure of model complexity in general for overparameterized models?

Short answers: YES.

Long answers: Rest of the talk.

Outline

Motivation and main punchlines

- Given $\mathcal{D} = \{(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}, 1 \le i \le n\}$, let $\widehat{\beta}_\lambda$ be ridge estimator:

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 / n + \lambda \|\beta\|_2^2$$

- The out-of-sample error of $\widehat{\beta}_\lambda$ is $y_0 - x_0^\top \widehat{\beta}_\lambda$ for a test point (x_0, y_0)

Key question: can we reliably estimate the entire out-of-sample error distribution and its linear and non-linear functionals in high dimensions?

We show, that under proportional asymptotics, almost surely:

1. the empirical distributions of re-weighted in-sample errors from leave-one-out and generalized cross-validation converge weakly to the out-of-sample error distribution, even when $\lambda = 0$
2. the plug-in estimators of these empirical distributions consistent for a broad class of linear and non-linear functionals of error distribution

Motivation and main punchlines

- Given $\mathcal{D} = \{(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}, 1 \leq i \leq n\}$, let $\widehat{\beta}_\lambda$ be ridge estimator:

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 / n + \lambda \|\beta\|_2^2$$

- The out-of-sample error of $\widehat{\beta}_\lambda$ is $y_0 - x_0^\top \widehat{\beta}_\lambda$ for a test point (x_0, y_0)

Key question: can we reliably estimate the entire out-of-sample error distribution and its linear and non-linear functionals in high dimensions?

We show, that under proportional asymptotics, almost surely:

1. the empirical distributions of re-weighted in-sample errors from leave-one-out and generalized cross-validation converge weakly to the out-of-sample error distribution, even when $\lambda = 0$
2. the plug-in estimators of these empirical distributions consistent for a broad class of linear and non-linear functionals of error distribution

Motivation and main punchlines

- Given $\mathcal{D} = \{(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}, 1 \leq i \leq n\}$, let $\widehat{\beta}_\lambda$ be ridge estimator:

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 / n + \lambda \|\beta\|_2^2$$

- The out-of-sample error of $\widehat{\beta}_\lambda$ is $y_0 - x_0^\top \widehat{\beta}_\lambda$ for a test point (x_0, y_0)

Key question: can we reliably estimate the entire out-of-sample error distribution and its linear and non-linear functionals in high dimensions?

We show, that under proportional asymptotics, almost surely:

1. the empirical distributions of re-weighted in-sample errors from leave-one-out and generalized cross-validation converge weakly to the out-of-sample error distribution, even when $\lambda = 0$
2. the plug-in estimators of these empirical distributions consistent for a broad class of linear and non-linear functionals of error distribution

Motivation and main punchlines

- Given $\mathcal{D} = \{(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}, 1 \le i \le n\}$, let $\widehat{\beta}_\lambda$ be ridge estimator:

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 / n + \lambda \|\beta\|_2^2$$

- The out-of-sample error of $\widehat{\beta}_\lambda$ is $y_0 - x_0^\top \widehat{\beta}_\lambda$ for a test point (x_0, y_0)

Key question: can we reliably estimate the entire out-of-sample error distribution and its linear and non-linear functionals in high dimensions?

We show, that under proportional asymptotics, almost surely:

1. the empirical distributions of re-weighted in-sample errors from leave-one-out and generalized cross-validation converge weakly to the out-of-sample error distribution, even when $\lambda = 0$
2. the plug-in estimators of these empirical distributions consistent for a broad class of linear and non-linear functionals of error distribution

Motivation and main punchlines

- Given $\mathcal{D} = \{(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}, 1 \leq i \leq n\}$, let $\widehat{\beta}_\lambda$ be ridge estimator:

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 / n + \lambda \|\beta\|_2^2$$

- The out-of-sample error of $\widehat{\beta}_\lambda$ is $y_0 - x_0^\top \widehat{\beta}_\lambda$ for a test point (x_0, y_0)

Key question: can we reliably estimate the entire out-of-sample error distribution and its linear and non-linear functionals in high dimensions?

We show, that under proportional asymptotics, almost surely:

1. the empirical distributions of re-weighted in-sample errors from leave-one-out and generalized cross-validation converge weakly to the out-of-sample error distribution, even when $\lambda = 0$
2. the plug-in estimators of these empirical distributions consistent for a broad class of linear and non-linear functionals of error distribution

Overview of high-dimensional ridge regression

- Let $X \in \mathbb{R}^{n \times p}$ denote feature matrix, $y \in \mathbb{R}^n$ denote response vector
- Let $\widehat{\beta}_\lambda \in \mathbb{R}^p$ denote the ridge estimator at regularization level λ:

$$\widehat{\beta}_\lambda := \arg\min_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2/n + \lambda\|\beta\|_2^2$$

  - if $\lambda > 0$, the problem is convex in β and has an explicit solution:

$$\widehat{\beta}_\lambda = (X^T X/n + \lambda I_p)^{-1} X^T y/n$$

  - for any $\lambda \in \mathbb{R}$, extend the solution using Moore-Penrose inverse:

$$\widehat{\beta}_\lambda = (X^T X/n + \lambda I_p)^{+} X^T y/n$$

  - when $\lambda = 0$, this reduces to least squares sol with minimum ℓ_2 norm; in particular, when $\text{rank}(X) = n \le p$, the solution interpolates data, i.e. $X\widehat{\beta} = y$, and has minimum ℓ_2 norm among all interpolators

Overview of high-dimensional ridge regression

- Let $X \in \mathbb{R}^{n \times p}$ denote feature matrix, $y \in \mathbb{R}^n$ denote response vector
- Let $\widehat{\beta}_\lambda \in \mathbb{R}^p$ denote the ridge estimator at regularization level λ:

$$\widehat{\beta}_\lambda := \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p} \ \|y - X\beta\|_2^2/n + \lambda\|\beta\|_2^2$$

  - if $\lambda > 0$, the problem is convex in β and has an explicit solution:

$$\widehat{\beta}_\lambda = (X^T X/n + \lambda I_p)^{-1} X^T y/n$$

  - for any $\lambda \in \mathbb{R}$, extend the solution using Moore-Penrose inverse:

$$\widehat{\beta}_\lambda = (X^T X/n + \lambda I_p)^{+} X^T y/n$$

  - when $\lambda = 0$, this reduces to least squares sol with minimum ℓ_2 norm; in particular, when $\operatorname{rank}(X) = n \le p$, the solution interpolates data, i.e. $X\widehat{\beta} = y$, and has minimum ℓ_2 norm among all interpolators

Overview of high-dimensional ridge regression

- Let $X \in \mathbb{R}^{n \times p}$ denote feature matrix, $y \in \mathbb{R}^n$ denote response vector
- Let $\widehat{\beta}_\lambda \in \mathbb{R}^p$ denote the ridge estimator at regularization level λ:

$$\widehat{\beta}_\lambda := \underset{\beta \in \mathbb{R}^p}{\arg\min} \ \|y - X\beta\|_2^2/n + \lambda\|\beta\|_2^2$$

  - if $\lambda > 0$, the problem is convex in β and has an explicit solution:

$$\widehat{\beta}_\lambda = (X^T X/n + \lambda I_p)^{-1} X^T y/n$$

  - for any $\lambda \in \mathbb{R}$, extend the solution using Moore-Penrose inverse:

$$\widehat{\beta}_\lambda = (X^T X/n + \lambda I_p)^{+} X^T y/n$$

  - when $\lambda = 0$, this reduces to least squares sol with minimum ℓ_2 norm; in particular, when $\text{rank}(X) = n \le p$, the solution interpolates data, i.e. $X\widehat{\beta} = y$, and has minimum ℓ_2 norm among all interpolators

Overview of high-dimensional ridge regression

- Let $X \in \mathbb{R}^{n \times p}$ denote feature matrix, $y \in \mathbb{R}^n$ denote response vector
- Let $\widehat{\beta}_\lambda \in \mathbb{R}^p$ denote the ridge estimator at regularization level λ:

$$\widehat{\beta}_\lambda := \arg\min_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2/n + \lambda\|\beta\|_2^2$$

  - if $\lambda > 0$, the problem is convex in β and has an explicit solution:

$$\widehat{\beta}_\lambda = (X^T X/n + \lambda I_p)^{-1} X^T y/n$$

  - for any $\lambda \in \mathbb{R}$, extend the solution using Moore-Penrose inverse:

$$\widehat{\beta}_\lambda = (X^T X/n + \lambda I_p)^{+} X^T y/n$$

  - when $\lambda = 0$, this reduces to least squares sol with minimum ℓ_2 norm; in particular, when $\text{rank}(X) = n \leq p$, the solution interpolates data, i.e. $X\widehat{\beta} = y$, and has minimum ℓ_2 norm among all interpolators

Overview of high-dimensional ridge regression

- Let $X \in \mathbb{R}^{n\times p}$ denote feature matrix, $y \in \mathbb{R}^n$ denote response vector
- Let $\widehat{\beta}_\lambda \in \mathbb{R}^p$ denote the ridge estimator at regularization level λ:

$$\widehat{\beta}_\lambda := \underset{\beta\in\mathbb{R}^p}{\arg\min}\ \|y - X\beta\|_2^2/n + \lambda\|\beta\|_2^2$$

  - if $\lambda > 0$, the problem is convex in β and has an explicit solution:

$$\widehat{\beta}_\lambda = (X^TX/n + \lambda I_p)^{-1}X^Ty/n$$

  - for any $\lambda \in \mathbb{R}$, extend the solution using Moore-Penrose inverse:

$$\widehat{\beta}_\lambda = (X^TX/n + \lambda I_p)^{+}X^Ty/n$$

  - when $\lambda = 0$, this reduces to least squares sol with minimum ℓ_2 norm; in particular, when $\text{rank}(X) = n \leq p$, the solution interpolates data, i.e. $X\widehat{\beta} = y$, and has minimum ℓ_2 norm among all interpolators

Out-of-sample error distribution and its functionals

- Let P_λ denote distribution of out-of-sample error of $\widehat{\beta}_\lambda$:

$$P_\lambda = \mathcal{L}\big(y_0 - x_0^\top \widehat{\beta}_\lambda \mid X, y\big),$$

  where (x_0, y_0) is sampled indep from the same training distribution
  - a random distribution (conditional on observed data X and y)
- Let ψ denote a functional such that $P \mapsto \psi(P) \in \mathbb{R}$:
  - Linear functional:

$$\psi(P_\lambda) = \int t(z)\, dP_\lambda(z) = \mathbb{E}\big[t(y_0 - x_0^\top \widehat{\beta}_\lambda) \mid X, y\big],$$

    where $t : \mathbb{R} \to \mathbb{R}$ is an error function (e.g., squared or absolute error)
  - Nonlinear functional:

$$\psi(P_\lambda) = \mathrm{Quantile}(P_\lambda; \tau) = \inf\{z : F_\lambda(z) \geq \tau\},$$

    where F_λ denotes the cumulative distribution function of P_λ

We construct estimators of P_λ and $\psi(P_\lambda)$ by suitably extending leave-one-out cross-validation and generalized cross-validation procedures.

Out-of-sample error distribution and its functionals

- Let P_λ denote distribution of out-of-sample error of $\widehat{\beta}_\lambda$:

$$P_\lambda = \mathcal{L}\big(y_0 - x_0^\top \widehat{\beta}_\lambda \mid X, y\big),$$

  where (x_0, y_0) is sampled indep from the same training distribution
  - a random distribution (conditional on observed data X and y)
- Let ψ denote a functional such that $P \mapsto \psi(P) \in \mathbb{R}$:
  - Linear functional:

$$\psi(P_\lambda) = \int t(z)\, dP_\lambda(z) = \mathbb{E}\big[t(y_0 - x_0^\top \widehat{\beta}_\lambda) \mid X, y\big],$$

    where $t : \mathbb{R} \to \mathbb{R}$ is an error function (e.g., squared or absolute error)
  - Nonlinear functional:

$$\psi(P_\lambda) = \mathrm{Quantile}(P_\lambda; \tau) = \inf\{z : F_\lambda(z) \geq \tau\},$$

    where F_λ denotes the cumulative distribution function of P_λ

We construct estimators of P_λ and $\psi(P_\lambda)$ by suitably extending leave-one-out cross-validation and generalized cross-validation procedures.

Out-of-sample error distribution and its functionals

- Let P_λ denote distribution of out-of-sample error of $\widehat{\beta}_\lambda$:

$$P_\lambda = \mathcal{L}\big(y_0 - x_0^\top \widehat{\beta}_\lambda \mid X, y\big),$$

  where (x_0, y_0) is sampled indep from the same training distribution
  - a random distribution (conditional on observed data X and y)
- Let ψ denote a functional such that $P \mapsto \psi(P) \in \mathbb{R}$:
  - Linear functional:

$$\psi(P_\lambda) = \int t(z)\, dP_\lambda(z) = \mathbb{E}\big[t(y_0 - x_0^\top \widehat{\beta}_\lambda) \mid X, y\big],$$

    where $t : \mathbb{R} \to \mathbb{R}$ is an error function (e.g., squared or absolute error)
  - Nonlinear functional:

$$\psi(P_\lambda) = \mathrm{Quantile}(P_\lambda; \tau) = \inf\{z : F_\lambda(z) \geq \tau\},$$

    where F_λ denotes the cumulative distribution function of P_λ

We construct estimators of P_λ and $\psi(P_\lambda)$ by suitably extending leave-one-out cross-validation and generalized cross-validation procedures.

Out-of-sample error distribution and its functionals

- Let P_λ denote distribution of out-of-sample error of $\widehat{\beta}_\lambda$:

$$P_\lambda = \mathcal{L}\big(y_0 - x_0^\top \widehat{\beta}_\lambda \mid X, y\big),$$

  where (x_0, y_0) is sampled indep from the same training distribution
  - a random distribution (conditional on observed data X and y)
- Let ψ denote a functional such that $P \mapsto \psi(P) \in \mathbb{R}$:
  - Linear functional:

$$\psi(P_\lambda) = \int t(z)\, dP_\lambda(z) = \mathbb{E}\big[t(y_0 - x_0^\top \widehat{\beta}_\lambda) \mid X, y\big],$$

    where $t : \mathbb{R} \to \mathbb{R}$ is an error function (e.g., squared or absolute error)
  - Nonlinear functional:

$$\psi(P_\lambda) = \mathrm{Quantile}(P_\lambda; \tau) = \inf\{z : F_\lambda(z) \geq \tau\},$$

    where F_λ denotes the cumulative distribution function of P_λ

We construct estimators of P_λ and $\psi(P_\lambda)$ by suitably extending leave-one-out cross-validation and generalized cross-validation procedures.

Out-of-sample error distribution and its functionals

- Let P_λ denote distribution of out-of-sample error of $\widehat{\beta}_\lambda$:

$$P_\lambda = \mathcal{L}\big(y_0 - x_0^\top \widehat{\beta}_\lambda \mid X, y\big),$$

  where (x_0, y_0) is sampled indep from the same training distribution
  - a random distribution (conditional on observed data X and y)
- Let ψ denote a functional such that $P \mapsto \psi(P) \in \mathbb{R}$:
  - Linear functional:

$$\psi(P_\lambda) = \int t(z)\, dP_\lambda(z) = \mathbb{E}\big[t(y_0 - x_0^\top \widehat{\beta}_\lambda) \mid X, y\big],$$

    where $t : \mathbb{R} \to \mathbb{R}$ is an error function (e.g., squared or absolute error)
  - Nonlinear functional:

$$\psi(P_\lambda) = \mathrm{Quantile}(P_\lambda; \tau) = \inf\{z : F_\lambda(z) \geq \tau\},$$

    where F_λ denotes the cumulative distribution function of P_λ

We construct estimators of P_λ and $\psi(P_\lambda)$ by suitably extending leave-one-out cross-validation and generalized cross-validation procedures.

Out-of-sample error distribution and its functionals

- Let P_λ denote distribution of out-of-sample error of $\widehat{\beta}_\lambda$:

$$P_\lambda = \mathcal{L}\big(y_0 - x_0^\top \widehat{\beta}_\lambda \mid X, y\big),$$

  where (x_0, y_0) is sampled indep from the same training distribution
  - a random distribution (conditional on observed data X and y)
- Let ψ denote a functional such that $P \mapsto \psi(P) \in \mathbb{R}$:
  - Linear functional:

$$\psi(P_\lambda) = \int t(z)\, dP_\lambda(z) = \mathbb{E}\big[t(y_0 - x_0^\top \widehat{\beta}_\lambda) \mid X, y\big],$$

    where $t : \mathbb{R} \to \mathbb{R}$ is an error function (e.g., squared or absolute error)
  - Nonlinear functional:

$$\psi(P_\lambda) = \mathrm{Quantile}(P_\lambda; \tau) = \inf\{z : F_\lambda(z) \geq \tau\},$$

    where F_λ denotes the cumulative distribution function of P_λ

We construct estimators of P_λ and $\psi(P_\lambda)$ by suitably extending leave-one-out cross-validation and generalized cross-validation procedures.

Standard leave-one-out and generalized cross-validation

- Leave-one-out cross-validation (LOOCV):
  - for every i, train on all data except (x_i, y_i), call the estimate $\widehat{\beta}_\lambda^{-i}$
  - compute test error on the i^{th} point and take average

$$\mathrm{loo}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - x_i^T\widehat{\beta}_\lambda^{-i}\right)^2$$

$$\overset{\text{(shortcut)}}{=} \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - x_i^T\widehat{\beta}_\lambda}{1-[L_\lambda]_{ii}}\right)^2$$

  where $L_\lambda = X(X^TX/n + \lambda I_p)^+X^T/n$ is the ridge smoothing matrix

- Generalized cross-validation (GCV)
  - same as leave-one-out shortcut but a single re-weighting

$$\mathrm{gcv}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - x_i^T\widehat{\beta}_\lambda}{1-\mathrm{tr}[L_\lambda]/n}\right)^2$$

- Standard LOOCV and GCV are consistent for the expected squared out-of-sample prediction error

Standard leave-one-out and generalized cross-validation

- Leave-one-out cross-validation (LOOCV):
  - for every i, train on all data except (x_i, y_i), call the estimate $\widehat{\beta}_\lambda^{-i}$
  - compute test error on the i^{th} point and take average

$$\mathrm{loo}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - x_i^T\widehat{\beta}_\lambda^{-i}\right)^2$$

$$\overset{\text{(shortcut)}}{=} \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - x_i^T\widehat{\beta}_\lambda}{1-[L_\lambda]_{ii}}\right)^2$$

where $L_\lambda = X(X^TX/n + \lambda I_p)^+X^T/n$ is the ridge smoothing matrix

- Generalized cross-validation (GCV)
  - same as leave-one-out shortcut but a single re-weighting

$$\mathrm{gcv}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - x_i^T\widehat{\beta}_\lambda}{1-\mathrm{tr}[L_\lambda]/n}\right)^2$$

- Standard LOOCV and GCV are consistent for the expected squared out-of-sample prediction error

Standard leave-one-out and generalized cross-validation

- Leave-one-out cross-validation (LOOCV):
  - for every i, train on all data except (x_i, y_i), call the estimate $\widehat{\beta}_\lambda^{-i}$
  - compute test error on the i^{th} point and take average

$$\mathrm{loo}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - x_i^T\widehat{\beta}_\lambda^{-i}\right)^2$$

$$\overset{\text{(shortcut)}}{=} \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - x_i^T\widehat{\beta}_\lambda}{1-[L_\lambda]_{ii}}\right)^2$$

where $L_\lambda = X(X^TX/n + \lambda I_p)^+X^T/n$ is the ridge smoothing matrix

- Generalized cross-validation (GCV)
  - same as leave-one-out shortcut but a single re-weighting

$$\mathrm{gcv}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - x_i^T\widehat{\beta}_\lambda}{1-\mathrm{tr}[L_\lambda]/n}\right)^2$$

- Standard LOOCV and GCV are consistent for the expected squared out-of-sample prediction error

Standard leave-one-out and generalized cross-validation

- Leave-one-out cross-validation (LOOCV):
  - for every i, train on all data except (x_i, y_i), call the estimate $\widehat{\beta}_\lambda^{-i}$
  - compute test error on the i^{th} point and take average

$$\text{loo}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - x_i^T\widehat{\beta}_\lambda^{-i}\right)^2$$

$$\overset{\text{(shortcut)}}{=} \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - x_i^T\widehat{\beta}_\lambda}{1-[L_\lambda]_{ii}}\right)^2$$

where $L_\lambda = X(X^TX/n + \lambda I_p)^+X^T/n$ is the ridge smoothing matrix

- Generalized cross-validation (GCV)
  - same as leave-one-out shortcut but a single re-weighting

$$\text{gcv}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - x_i^T\widehat{\beta}_\lambda}{1-\text{tr}[L_\lambda]/n}\right)^2$$

- Standard LOOCV and GCV are consistent for the expected squared out-of-sample prediction error

Standard leave-one-out and generalized cross-validation

- Leave-one-out cross-validation (LOOCV):
  - for every i, train on all data except (x_i, y_i), call the estimate $\widehat{\beta}_\lambda^{-i}$
  - compute test error on the i^{th} point and take average

$$\mathrm{loo}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - x_i^T\widehat{\beta}_\lambda^{-i}\right)^2$$

$$\overset{\text{(shortcut)}}{=} \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - x_i^T\widehat{\beta}_\lambda}{1-[L_\lambda]_{ii}}\right)^2$$

where $L_\lambda = X(X^TX/n + \lambda I_p)^{+}X^T/n$ is the ridge smoothing matrix

- Generalized cross-validation (GCV)
  - same as leave-one-out shortcut but a single re-weighting

$$\mathrm{gcv}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - x_i^T\widehat{\beta}_\lambda}{1-\mathrm{tr}[L_\lambda]/n}\right)^2$$

- Standard LOOCV and GCV are consistent for the expected squared out-of-sample prediction error

Proposed estimators

Natural estimators for P_λ and $\psi(P_\lambda)$ building off from GCV and LOOCV.

- Empirical distributions of the GCV, LOO re-weighted errors:

$$\widehat{P}_\lambda^{\mathrm{gcv}} = \frac{1}{n}\sum_{i=1}^n \delta\left(\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - \mathrm{tr}[L_\lambda]/n}\right) \quad \text{and} \quad \widehat{P}_\lambda^{\mathrm{loo}} = \frac{1}{n}\sum_{i=1}^n \delta\left(\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - [L_\lambda]_{ii}}\right)$$

- When $\widehat{\beta}_\lambda$ is an interpolator, i.e. $L_\lambda = I_n$, both estimates are "0/0"[2]; we then define the estimates as their respective limits as $\lambda \to 0$:

$$\widehat{P}_0^{\mathrm{gcv}} = \frac{1}{n}\sum_{i=1}^n \delta\left(\frac{[(XX^\top)^\dagger y]_i}{\mathrm{tr}[(XX^\top)^\dagger]/n}\right) \quad \text{and} \quad \widehat{P}_0^{\mathrm{loo}} = \frac{1}{n}\sum_{i=1}^n \delta\left(\frac{[(XX^\top)^\dagger y]_i}{[(XX^\top)^\dagger]_{ii}}\right)$$

- Plug-in GCV and LOO estimators:

$$\widehat{\psi}_\lambda^{\mathrm{gcv}} = \psi(\widehat{P}_\lambda^{\mathrm{gcv}}) \quad \text{and} \quad \widehat{\psi}_\lambda^{\mathrm{loo}} = \psi(\widehat{P}_\lambda^{\mathrm{gcv}})$$

[2] The idea of analytic continuation at $\lambda = 0$ is from Hastie, Montanari, Rosset, Tibshirani, 2019: "Surprises in high-dimensional ridgeless least squares interpolation"

Proposed estimators

Natural estimators for P_λ and $\psi(P_\lambda)$ building off from GCV and LOOCV.

- Empirical distributions of the GCV, LOO re-weighted errors:

$$\widehat{P}_\lambda^{\mathrm{gcv}} = \frac{1}{n}\sum_{i=1}^{n} \delta\left(\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - \mathrm{tr}[L_\lambda]/n}\right) \quad \text{and} \quad \widehat{P}_\lambda^{\mathrm{loo}} = \frac{1}{n}\sum_{i=1}^{n} \delta\left(\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - [L_\lambda]_{ii}}\right)$$

- When $\widehat{\beta}_\lambda$ is an interpolator, i.e. $L_\lambda = I_n$, both estimates are "0/0"[2]; we then define the estimates as their respective limits as $\lambda \to 0$:

$$\widehat{P}_0^{\mathrm{gcv}} = \frac{1}{n}\sum_{i=1}^{n} \delta\left(\frac{[(XX^\top)^\dagger y]_i}{\mathrm{tr}[(XX^\top)^\dagger]/n}\right) \quad \text{and} \quad \widehat{P}_0^{\mathrm{loo}} = \frac{1}{n}\sum_{i=1}^{n} \delta\left(\frac{[(XX^\top)^\dagger y]_i}{[(XX^\top)^\dagger]_{ii}}\right)$$

- Plug-in GCV and LOO estimators:

$$\widehat{\psi}_\lambda^{\mathrm{gcv}} = \psi(\widehat{P}_\lambda^{\mathrm{gcv}}) \quad \text{and} \quad \widehat{\psi}_\lambda^{\mathrm{loo}} = \psi(\widehat{P}_\lambda^{\mathrm{gcv}})$$

[2] The idea of analytic continuation at $\lambda = 0$ is from Hastie, Montanari, Rosset, Tibshirani, 2019: "Surprises in high-dimensional ridgeless least squares interpolation"

Proposed estimators

Natural estimators for P_λ and $\psi(P_\lambda)$ building off from GCV and LOOCV.

- Empirical distributions of the GCV, LOO re-weighted errors:

$$\widehat{P}_\lambda^{\mathrm{gcv}} = \frac{1}{n}\sum_{i=1}^{n} \delta\left(\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - \mathrm{tr}[L_\lambda]/n}\right) \quad \text{and} \quad \widehat{P}_\lambda^{\mathrm{loo}} = \frac{1}{n}\sum_{i=1}^{n} \delta\left(\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - [L_\lambda]_{ii}}\right)$$

- When $\widehat{\beta}_\lambda$ is an interpolator, i.e. $L_\lambda = I_n$, both estimates are "0/0"[2]; we then define the estimates as their respective limits as $\lambda \to 0$:

$$\widehat{P}_0^{\mathrm{gcv}} = \frac{1}{n}\sum_{i=1}^{n} \delta\left(\frac{[(XX^\top)^\dagger y]_i}{\mathrm{tr}[(XX^\top)^\dagger]/n}\right) \quad \text{and} \quad \widehat{P}_0^{\mathrm{loo}} = \frac{1}{n}\sum_{i=1}^{n} \delta\left(\frac{[(XX^\top)^\dagger y]_i}{[(XX^\top)^\dagger]_{ii}}\right)$$

- Plug-in GCV and LOO estimators:

$$\widehat{\psi}_\lambda^{\mathrm{gcv}} = \psi(\widehat{P}_\lambda^{\mathrm{gcv}}) \quad \text{and} \quad \widehat{\psi}_\lambda^{\mathrm{loo}} = \psi(\widehat{P}_\lambda^{\mathrm{gcv}})$$

[2] The idea of analytic continuation at $\lambda = 0$ is from Hastie, Montanari, Rosset, Tibshirani, 2019: "Surprises in high-dimensional ridgeless least squares interpolation"

Proposed estimators

Natural estimators for P_λ and $\psi(P_\lambda)$ building off from GCV and LOOCV.

- Empirical distributions of the GCV, LOO re-weighted errors:

$$\widehat{P}_\lambda^{\mathrm{gcv}} = \frac{1}{n}\sum_{i=1}^{n} \delta\left(\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - \mathrm{tr}[L_\lambda]/n}\right) \quad \text{and} \quad \widehat{P}_\lambda^{\mathrm{loo}} = \frac{1}{n}\sum_{i=1}^{n} \delta\left(\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - [L_\lambda]_{ii}}\right)$$

- When $\widehat{\beta}_\lambda$ is an interpolator, i.e. $L_\lambda = I_n$, both estimates are "0/0"[2]; we then define the estimates as their respective limits as $\lambda \to 0$:

$$\widehat{P}_0^{\mathrm{gcv}} = \frac{1}{n}\sum_{i=1}^{n} \delta\left(\frac{[(XX^\top)^\dagger y]_i}{\mathrm{tr}[(XX^\top)^\dagger]/n}\right) \quad \text{and} \quad \widehat{P}_0^{\mathrm{loo}} = \frac{1}{n}\sum_{i=1}^{n} \delta\left(\frac{[(XX^\top)^\dagger y]_i}{[(XX^\top)^\dagger]_{ii}}\right)$$

- Plug-in GCV and LOO estimators:

$$\widehat{\psi}_\lambda^{\mathrm{gcv}} = \psi(\widehat{P}_\lambda^{\mathrm{gcv}}) \quad \text{and} \quad \widehat{\psi}_\lambda^{\mathrm{loo}} = \psi(\widehat{P}_\lambda^{\mathrm{gcv}})$$

[2] The idea of analytic continuation at $\lambda = 0$ is from Hastie, Montanari, Rosset, Tibshirani, 2019: "Surprises in high-dimensional ridgeless least squares interpolation"

Distribution estimation

Theorem. Under i.i.d. sampling of (x_i, y_i), $i = 1, \ldots, n$ with

1. feature x_i decomposable into $x_i = \Sigma^{1/2} z_i$ where z_i contains i.i.d. entries with mean 0, variance 1, and finite 4+ moment, and spectrum of Σ is uniformly away from $r_{\min} > 0$ and $r_{\max} < \infty$,
2. response y_i with bounded 4+ moment,

as $n, p \to \infty$ such that $p/n \to \gamma \in (0, \infty)$, almost surely, for each $\lambda > \lambda_{\min} := -(1 - \sqrt{\gamma})^2 r_{\min} \leq 0$,

$$\widehat{P}_\lambda^{\mathrm{gcv}} \xrightarrow{\mathrm{d}} P_\lambda, \quad \text{and} \quad \widehat{P}_\lambda^{\mathrm{loo}} \xrightarrow{\mathrm{d}} P_\lambda.$$

Remarks:

- Almost sure convergence with respect to the training data
- The regression function does not need to be linear in x
- Amazingly, this results also holds when $\lambda = 0$ (min-norm estimator)

Distribution estimation

Theorem. Under i.i.d. sampling of (x_i, y_i), $i = 1, \ldots, n$ with

1. feature x_i decomposable into $x_i = \Sigma^{1/2} z_i$ where z_i contains i.i.d. entries with mean 0, variance 1, and finite 4+ moment, and spectrum of Σ is uniformly away from $r_{\min} > 0$ and $r_{\max} < \infty$,
2. response y_i with bounded 4+ moment,

as $n, p \to \infty$ such that $p/n \to \gamma \in (0, \infty)$, almost surely, for each $\lambda > \lambda_{\min} := -(1 - \sqrt{\gamma})^2 r_{\min} \leq 0$,

$$\widehat{P}_\lambda^{\mathrm{gcv}} \xrightarrow{\mathrm{d}} P_\lambda, \quad \text{and} \quad \widehat{P}_\lambda^{\mathrm{loo}} \xrightarrow{\mathrm{d}} P_\lambda.$$

Remarks:

- Almost sure convergence with respect to the training data
- The regression function does not need to be linear in x
- Amazingly, this results also holds when $\lambda = 0$ (min-norm estimator)

Distribution estimation

Theorem. Under i.i.d. sampling of (x_i, y_i), $i = 1, \ldots, n$ with

1. feature x_i decomposable into $x_i = \Sigma^{1/2} z_i$ where z_i contains i.i.d. entries with mean 0, variance 1, and finite 4+ moment, and spectrum of Σ is uniformly away from $r_{\min} > 0$ and $r_{\max} < \infty$,
2. response y_i with bounded 4+ moment,

as $n, p \to \infty$ such that $p/n \to \gamma \in (0, \infty)$, almost surely, for each $\lambda > \lambda_{\min} := -(1 - \sqrt{\gamma})^2 r_{\min} \leq 0$,

$$\widehat{P}_\lambda^{\mathrm{gcv}} \xrightarrow{\mathrm{d}} P_\lambda, \quad \text{and} \quad \widehat{P}_\lambda^{\mathrm{loo}} \xrightarrow{\mathrm{d}} P_\lambda.$$

Remarks:

- Almost sure convergence with respect to the training data
- The regression function does not need to be linear in x
- Amazingly, this results also holds when $\lambda = 0$ (min-norm estimator)

Distribution estimation

Theorem. Under i.i.d. sampling of (x_i, y_i), $i = 1, \ldots, n$ with

1. feature x_i decomposable into $x_i = \Sigma^{1/2} z_i$ where z_i contains i.i.d. entries with mean 0, variance 1, and finite 4+ moment, and spectrum of Σ is uniformly away from $r_{\min} > 0$ and $r_{\max} < \infty$,
2. response y_i with bounded 4+ moment,

as $n, p \to \infty$ such that $p/n \to \gamma \in (0, \infty)$, almost surely, for each $\lambda > \lambda_{\min} := -(1 - \sqrt{\gamma})^2 r_{\min} \leq 0$,

$$\widehat{P}_\lambda^{\text{gcv}} \xrightarrow{\text{d}} P_\lambda, \quad \text{and} \quad \widehat{P}_\lambda^{\text{loo}} \xrightarrow{\text{d}} P_\lambda.$$

Remarks:

- Almost sure convergence with respect to the training data
- The regression function does not need to be linear in x
- Amazingly, this results also holds when $\lambda = 0$ (min-norm estimator)

Distribution estimation

Theorem. Under i.i.d. sampling of (x_i, y_i), $i = 1, \ldots, n$ with

1. feature x_i decomposable into $x_i = \Sigma^{1/2} z_i$ where z_i contains i.i.d. entries with mean 0, variance 1, and finite 4+ moment, and spectrum of Σ is uniformly away from $r_{\min} > 0$ and $r_{\max} < \infty$,
2. response y_i with bounded 4+ moment,

as $n, p \to \infty$ such that $p/n \to \gamma \in (0, \infty)$, almost surely, for each $\lambda > \lambda_{\min} := -(1 - \sqrt{\gamma})^2 r_{\min} \leq 0$,

$$\widehat{P}_\lambda^{\mathrm{gcv}} \xrightarrow{\mathrm{d}} P_\lambda, \quad \text{and} \quad \widehat{P}_\lambda^{\mathrm{loo}} \xrightarrow{\mathrm{d}} P_\lambda.$$

Remarks:

- Almost sure convergence with respect to the training data
- The regression function does not need to be linear in x
- Amazingly, this results also holds when $\lambda = 0$ (min-norm estimator)

Distribution estimation

Theorem. Under i.i.d. sampling of (x_i, y_i), $i = 1, \dots, n$ with

1. feature x_i decomposable into $x_i = \Sigma^{1/2} z_i$ where z_i contains i.i.d. entries with mean 0, variance 1, and finite $4+$ moment, and spectrum of Σ is uniformly away from $r_{\min} > 0$ and $r_{\max} < \infty$,
2. response y_i with bounded $4+$ moment,

as $n, p \to \infty$ such that $p/n \to \gamma \in (0, \infty)$, almost surely, for each $\lambda > \lambda_{\min} := -(1 - \sqrt{\gamma})^2 r_{\min} \leq 0$,

$$\widehat{P}_\lambda^{\mathrm{gcv}} \xrightarrow{\mathrm{d}} P_\lambda, \quad \text{and} \quad \widehat{P}_\lambda^{\mathrm{loo}} \xrightarrow{\mathrm{d}} P_\lambda.$$

Remarks:

- Almost sure convergence with respect to the training data
- The regression function does not need to be linear in x
- Amazingly, this results also holds when $\lambda = 0$ (min-norm estimator)

Distribution estimation

Theorem. Under i.i.d. sampling of (x_i, y_i), $i = 1, \ldots, n$ with

1. feature x_i decomposable into $x_i = \Sigma^{1/2} z_i$ where z_i contains i.i.d. entries with mean 0, variance 1, and finite 4+ moment, and spectrum of Σ is uniformly away from $r_{\min} > 0$ and $r_{\max} < \infty$,
2. response y_i with bounded 4+ moment,

as $n, p \to \infty$ such that $p/n \to \gamma \in (0, \infty)$, almost surely, for each $\lambda > \lambda_{\min} := -(1 - \sqrt{\gamma})^2 r_{\min} \leq 0$,

$$\widehat{P}_\lambda^{\mathrm{gcv}} \xrightarrow{\mathrm{d}} P_\lambda, \quad \text{and} \quad \widehat{P}_\lambda^{\mathrm{loo}} \xrightarrow{\mathrm{d}} P_\lambda.$$

Remarks:

- Almost sure convergence with respect to the training data
- The regression function does not need to be linear in x
- Amazingly, this results also holds when $\lambda = 0$ (min-norm estimator)

Distribution estimation

Theorem. Under i.i.d. sampling of (x_i, y_i), $i = 1, \ldots, n$ with

1. feature x_i decomposable into $x_i = \Sigma^{1/2} z_i$ where z_i contains i.i.d. entries with mean 0, variance 1, and finite 4+ moment, and spectrum of Σ is uniformly away from $r_{\min} > 0$ and $r_{\max} < \infty$,
2. response y_i with bounded 4+ moment,

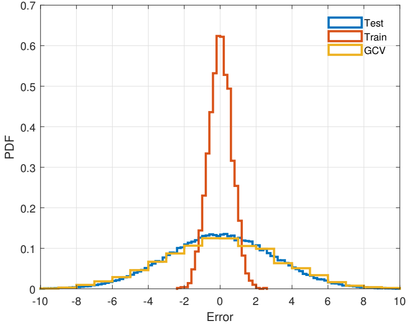
as $n, p \to \infty$ such that $p/n \to \gamma \in (0, \infty)$, almost surely, for each $\lambda > \lambda_{\min} := -(1 - \sqrt{\gamma})^2 r_{\min} \le 0$,

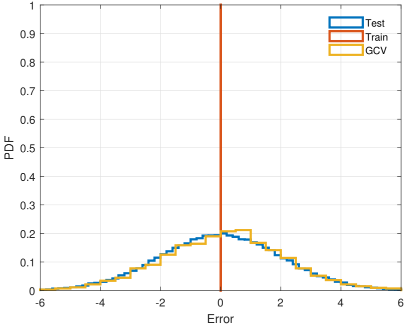
$$\widehat{P}_\lambda^{\mathrm{gcv}} \xrightarrow{\mathrm{d}} P_\lambda, \quad \text{and} \quad \widehat{P}_\lambda^{\mathrm{loo}} \xrightarrow{\mathrm{d}} P_\lambda.$$

Remarks:

- Almost sure convergence with respect to the training data
- The regression function does not need to be linear in x
- Amazingly, this results also holds when $\lambda = 0$ (min-norm estimator)

Distribution estimation: illustration

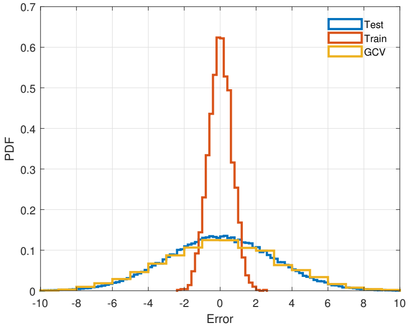


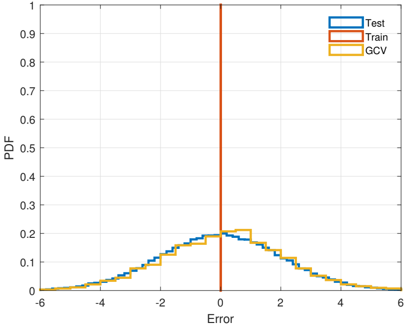


- $n = 2500$, $p = 2000$, $p/n = 0.8$
- $\lambda = 0$, i.e., least squares

- $n = 2500$, $p = 5000$, $p/n = 2$
- $\lambda = 0$, i.e., min-norm estimator, zero in-sample errors

Distribution estimation: illustration

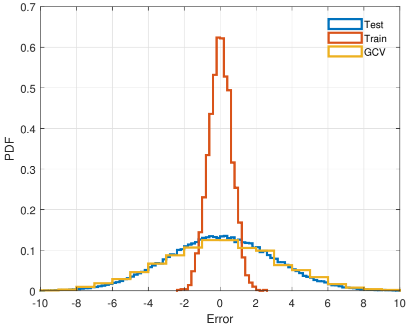


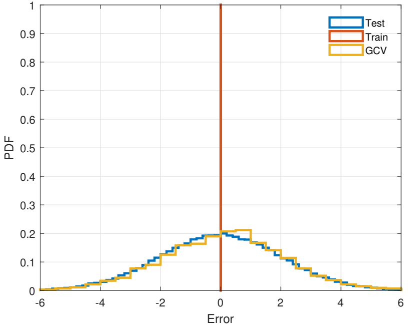


- $n = 2500,\ p = 2000,\ p/n = 0.8$
- $\lambda = 0$, i.e., least squares

- $n = 2500,\ p = 5000,\ p/n = 2$
- $\lambda = 0$, i.e., min-norm estimator, zero in-sample errors

Distribution estimation: illustration

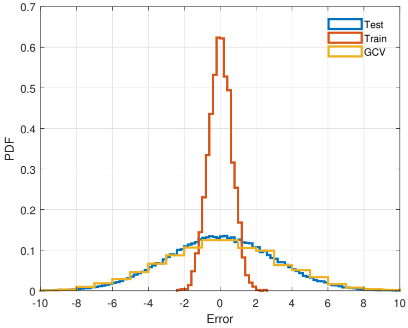


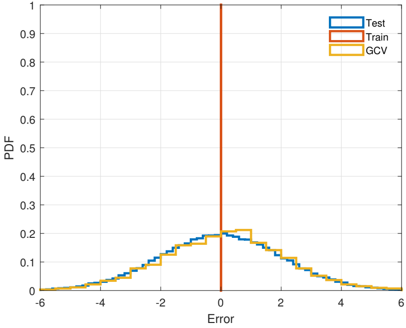


- $n = 2500,\ p = 2000,\ p/n = 0.8$
- $\lambda = 0$, i.e., least squares

- $n = 2500,\ p = 5000,\ p/n = 2$
- $\lambda = 0$, i.e., min-norm estimator, zero in-sample errors

Distribution estimation: illustration

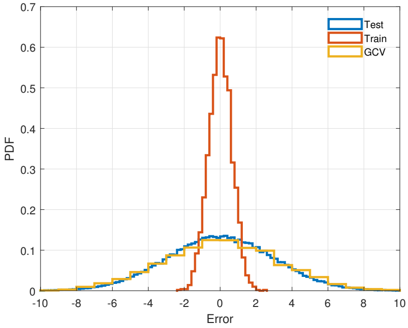


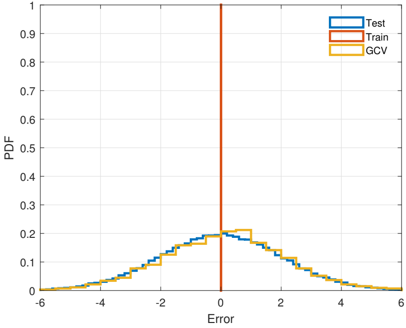


- $n = 2500,\ p = 2000,\ p/n = 0.8$
- $\lambda = 0$, i.e., least squares

- $n = 2500,\ p = 5000,\ p/n = 2$
- $\lambda = 0$, i.e., min-norm estimator, zero in-sample errors

Distribution estimation: illustration





- $n = 2500$, $p = 2000$, $p/n = 0.8$
- $\lambda = 0$, i.e., least squares

- $n = 2500$, $p = 5000$, $p/n = 2$
- $\lambda = 0$, i.e., min-norm estimator, zero in-sample errors

Linear functional estimation (pointwise in λ)

- Let T_λ be a linear functional of the out-of-sample error distribution:

$$T_\lambda = \mathbb{E}\big[t(y_0 - x_0^T \widehat{\beta}_\lambda) \mid X, y\big]$$

- Let $\widehat{T}_\lambda^{\mathrm{gcv}}$ and $\widehat{T}_\lambda^{\mathrm{loo}}$ be plug-in estimators from GCV and LOOCV:

$$\widehat{T}_\lambda^{\mathrm{gcv}} = \frac{1}{n}\sum_{i=1}^{n} t\left(\frac{y_i - x_i^T\widehat{\beta}_\lambda}{1 - \mathrm{tr}[L_\lambda]/n}\right) \quad \text{and} \quad \widehat{T}_\lambda^{\mathrm{loo}} = \frac{1}{n}\sum_{i=1}^{n} t\left(\frac{y_i - x_i^T\widehat{\beta}_\lambda}{1 - [L_\lambda]_{ii}}\right)$$

Theorem. For error functions $t : \mathbb{R} \to \mathbb{R}$

1. that are continuous
2. have quadratic growth, i.e., there exist constants $a, b, c > 0$ such that $|t(z)| \le az^2 + b|z| + c$ for any $z \in \mathbb{R}$,

as $n, p \to \infty$ with $p/n \to \gamma \in (0, \infty)$, for $\lambda > \lambda_{\min} := -(1-\sqrt{\gamma})^2 r_{\min}$,

$$\widehat{T}_\lambda^{\mathrm{gcv}} \xrightarrow{\text{a.s.}} T_\lambda, \quad \text{and} \quad \widehat{T}_\lambda^{\mathrm{loo}} \xrightarrow{\text{a.s.}} T_\lambda.$$

Linear functional estimation (pointwise in λ)

- Let T_λ be a linear functional of the out-of-sample error distribution:

$$T_\lambda = \mathbb{E}\big[t(y_0 - x_0^T \widehat{\beta}_\lambda) \mid X, y\big]$$

- Let $\widehat{T}_\lambda^{\mathrm{gcv}}$ and $\widehat{T}_\lambda^{\mathrm{loo}}$ be plug-in estimators from GCV and LOOCV:

$$\widehat{T}_\lambda^{\mathrm{gcv}} = \frac{1}{n}\sum_{i=1}^{n} t\left(\frac{y_i - x_i^T\widehat{\beta}_\lambda}{1 - \mathrm{tr}[L_\lambda]/n}\right) \quad \text{and} \quad \widehat{T}_\lambda^{\mathrm{loo}} = \frac{1}{n}\sum_{i=1}^{n} t\left(\frac{y_i - x_i^T\widehat{\beta}_\lambda}{1 - [L_\lambda]_{ii}}\right)$$

Theorem. For error functions $t : \mathbb{R} \to \mathbb{R}$

1. that are continuous
2. have quadratic growth, i.e., there exist constants $a, b, c > 0$ such that $|t(z)| \leq az^2 + b|z| + c$ for any $z \in \mathbb{R}$,

as $n, p \to \infty$ with $p/n \to \gamma \in (0, \infty)$, for $\lambda > \lambda_{\min} := -(1 - \sqrt{\gamma})^2 r_{\min}$,

$$\widehat{T}_\lambda^{\mathrm{gcv}} \xrightarrow{\text{a.s.}} T_\lambda, \quad \text{and} \quad \widehat{T}_\lambda^{\mathrm{loo}} \xrightarrow{\text{a.s.}} T_\lambda.$$

Linear functional estimation (pointwise in λ)

- Let T_λ be a linear functional of the out-of-sample error distribution:

$$T_\lambda = \mathbb{E}\big[t(y_0 - x_0^T \widehat{\beta}_\lambda) \mid X, y\big]$$

- Let $\widehat{T}_\lambda^{\mathrm{gcv}}$ and $\widehat{T}_\lambda^{\mathrm{loo}}$ be plug-in estimators from GCV and LOOCV:

$$\widehat{T}_\lambda^{\mathrm{gcv}} = \frac{1}{n}\sum_{i=1}^{n} t\left(\frac{y_i - x_i^T\widehat{\beta}_\lambda}{1 - \mathrm{tr}[L_\lambda]/n}\right) \quad \text{and} \quad \widehat{T}_\lambda^{\mathrm{loo}} = \frac{1}{n}\sum_{i=1}^{n} t\left(\frac{y_i - x_i^T\widehat{\beta}_\lambda}{1 - [L_\lambda]_{ii}}\right)$$

Theorem. For error functions $t : \mathbb{R} \to \mathbb{R}$

1. that are continuous
2. have quadratic growth, i.e., there exist constants $a, b, c > 0$ such that $|t(z)| \le az^2 + b|z| + c$ for any $z \in \mathbb{R}$,

as $n, p \to \infty$ with $p/n \to \gamma \in (0, \infty)$, for $\lambda > \lambda_{\min} := -(1 - \sqrt{\gamma})^2 r_{\min}$,

$$\widehat{T}_\lambda^{\mathrm{gcv}} \xrightarrow{\text{a.s.}} T_\lambda, \quad \text{and} \quad \widehat{T}_\lambda^{\mathrm{loo}} \xrightarrow{\text{a.s.}} T_\lambda.$$

Linear functional estimation (pointwise in λ)

- Let T_λ be a linear functional of the out-of-sample error distribution:

$$T_\lambda = \mathbb{E}\big[t(y_0 - x_0^T \widehat{\beta}_\lambda) \mid X, y\big]$$

- Let $\widehat{T}_\lambda^{\mathrm{gcv}}$ and $\widehat{T}_\lambda^{\mathrm{loo}}$ be plug-in estimators from GCV and LOOCV:

$$\widehat{T}_\lambda^{\mathrm{gcv}} = \frac{1}{n} \sum_{i=1}^{n} t\left(\frac{y_i - x_i^T \widehat{\beta}_\lambda}{1 - \mathrm{tr}[L_\lambda]/n}\right) \quad \text{and} \quad \widehat{T}_\lambda^{\mathrm{loo}} = \frac{1}{n} \sum_{i=1}^{n} t\left(\frac{y_i - x_i^T \widehat{\beta}_\lambda}{1 - [L_\lambda]_{ii}}\right)$$

Theorem. For error functions $t : \mathbb{R} \to \mathbb{R}$

1. that are continuous
2. have quadratic growth, i.e., there exist constants $a, b, c > 0$ such that $|t(z)| \le az^2 + b|z| + c$ for any $z \in \mathbb{R}$,

as $n, p \to \infty$ with $p/n \to \gamma \in (0, \infty)$, for $\lambda > \lambda_{\min} := -(1 - \sqrt{\gamma})^2 r_{\min}$,

$$\widehat{T}_\lambda^{\mathrm{gcv}} \xrightarrow{\text{a.s.}} T_\lambda, \quad \text{and} \quad \widehat{T}_\lambda^{\mathrm{loo}} \xrightarrow{\text{a.s.}} T_\lambda.$$

Linear functional estimation (pointwise in λ)

- Let T_λ be a linear functional of the out-of-sample error distribution:

$$T_\lambda = \mathbb{E}\big[t(y_0 - x_0^T \widehat{\beta}_\lambda) \mid X, y\big]$$

- Let $\widehat{T}_\lambda^{\mathrm{gcv}}$ and $\widehat{T}_\lambda^{\mathrm{loo}}$ be plug-in estimators from GCV and LOOCV:

$$\widehat{T}_\lambda^{\mathrm{gcv}} = \frac{1}{n}\sum_{i=1}^{n} t\left(\frac{y_i - x_i^T\widehat{\beta}_\lambda}{1 - \mathrm{tr}[L_\lambda]/n}\right) \quad \text{and} \quad \widehat{T}_\lambda^{\mathrm{loo}} = \frac{1}{n}\sum_{i=1}^{n} t\left(\frac{y_i - x_i^T\widehat{\beta}_\lambda}{1 - [L_\lambda]_{ii}}\right)$$

Theorem. For error functions $t : \mathbb{R} \to \mathbb{R}$

1. that are continuous
2. have quadratic growth, i.e., there exist constants $a, b, c > 0$ such that $|t(z)| \le az^2 + b|z| + c$ for any $z \in \mathbb{R}$,

as $n, p \to \infty$ with $p/n \to \gamma \in (0, \infty)$, for $\lambda > \lambda_{\min} := -(1 - \sqrt{\gamma})^2 r_{\min}$,

$$\widehat{T}_\lambda^{\mathrm{gcv}} \xrightarrow{\text{a.s.}} T_\lambda, \quad \text{and} \quad \widehat{T}_\lambda^{\mathrm{loo}} \xrightarrow{\text{a.s.}} T_\lambda.$$

Linear functional estimation (pointwise in λ)

- Let T_λ be a linear functional of the out-of-sample error distribution:

$$T_\lambda = \mathbb{E}\big[t(y_0 - x_0^T \widehat{\beta}_\lambda) \mid X, y\big]$$

- Let $\widehat{T}_\lambda^{\mathrm{gcv}}$ and $\widehat{T}_\lambda^{\mathrm{loo}}$ be plug-in estimators from GCV and LOOCV:

$$\widehat{T}_\lambda^{\mathrm{gcv}} = \frac{1}{n} \sum_{i=1}^{n} t\left(\frac{y_i - x_i^T \widehat{\beta}_\lambda}{1 - \mathrm{tr}[L_\lambda]/n}\right) \quad \text{and} \quad \widehat{T}_\lambda^{\mathrm{loo}} = \frac{1}{n} \sum_{i=1}^{n} t\left(\frac{y_i - x_i^T \widehat{\beta}_\lambda}{1 - [L_\lambda]_{ii}}\right)$$

Theorem. For error functions $t : \mathbb{R} \to \mathbb{R}$

1. that are continuous
2. have quadratic growth, i.e., there exist constants $a, b, c > 0$ such that $|t(z)| \le az^2 + b|z| + c$ for any $z \in \mathbb{R}$,

as $n, p \to \infty$ with $p/n \to \gamma \in (0, \infty)$, for $\lambda > \lambda_{\mathrm{min}} := -(1 - \sqrt{\gamma})^2 r_{\mathrm{min}}$,

$$\widehat{T}_\lambda^{\mathrm{gcv}} \xrightarrow{\mathrm{a.s.}} T_\lambda, \quad \text{and} \quad \widehat{T}_\lambda^{\mathrm{loo}} \xrightarrow{\mathrm{a.s.}} T_\lambda.$$

Linear functional estimation (uniform in λ)

Theorem. For error functions $t : \mathbb{R} \to \mathbb{R}$

1. that are differentiable
2. have derivative with linear growth rate, i.e., there exist constants $g, h > 0$ such that $|t'(z)| \leq g|z| + h$ for any $z \in \mathbb{R}$

as $n, p \to \infty$ with $p/n \to \gamma \in (0, \infty)$ for any compact set $\Lambda \subseteq (\lambda_{\min}, \infty)$,

$$\sup_{\lambda \in \Lambda} \left| \widehat{T}_\lambda^{\text{gcv}} - T_\lambda \right| \xrightarrow{\text{a.s.}} 0, \quad \text{and} \quad \sup_{\lambda \in \Lambda} \left| \widehat{T}_\lambda^{\text{loo}} - T_\lambda \right| \xrightarrow{\text{a.s.}} 0.$$

Remarks:

- Special case of $t(r) = r^2$ exploits bias-variance decomposition
- No bias-variance decomposition for general error functions and result requires a different proof technique via leave-one-out arguments
- Using uniformity arguments, the result can be extended for non-linear variational functionals

Linear functional estimation (uniform in λ)

Theorem. For error functions $t : \mathbb{R} \to \mathbb{R}$

1. that are differentiable
2. have derivative with linear growth rate, i.e., there exist constants $g, h > 0$ such that $|t'(z)| \leq g|z| + h$ for any $z \in \mathbb{R}$

as $n, p \to \infty$ with $p/n \to \gamma \in (0, \infty)$ for any compact set $\Lambda \subseteq (\lambda_{\min}, \infty)$,

$$\sup_{\lambda \in \Lambda} \left| \widehat{T}_\lambda^{\mathrm{gcv}} - T_\lambda \right| \xrightarrow{\text{a.s.}} 0, \quad \text{and} \quad \sup_{\lambda \in \Lambda} \left| \widehat{T}_\lambda^{\mathrm{loo}} - T_\lambda \right| \xrightarrow{\text{a.s.}} 0.$$

Remarks:

- Special case of $t(r) = r^2$ exploits bias-variance decomposition
- No bias-variance decomposition for general error functions and result requires a different proof technique via leave-one-out arguments
- Using uniformity arguments, the result can be extended for non-linear variational functionals

Linear functional estimation (uniform in λ)

Theorem. For error functions $t : \mathbb{R} \to \mathbb{R}$

1. that are differentiable
2. have derivative with linear growth rate, i.e., there exist constants $g, h > 0$ such that $|t'(z)| \leq g|z| + h$ for any $z \in \mathbb{R}$

as $n, p \to \infty$ with $p/n \to \gamma \in (0, \infty)$ for any compact set $\Lambda \subseteq (\lambda_{\min}, \infty)$,

$$\sup_{\lambda \in \Lambda} \left| \widehat{T}_\lambda^{\mathrm{gcv}} - T_\lambda \right| \xrightarrow{\text{a.s.}} 0, \quad \text{and} \quad \sup_{\lambda \in \Lambda} \left| \widehat{T}_\lambda^{\mathrm{loo}} - T_\lambda \right| \xrightarrow{\text{a.s.}} 0.$$

Remarks:

- Special case of $t(r) = r^2$ exploits bias-variance decomposition
- No bias-variance decomposition for general error functions and result requires a different proof technique via leave-one-out arguments
- Using uniformity arguments, the result can be extended for non-linear variational functionals

Linear functional estimation (uniform in λ)

Theorem. For error functions $t : \mathbb{R} \to \mathbb{R}$

1. that are differentiable
2. have derivative with linear growth rate, i.e., there exist constants $g, h > 0$ such that $|t'(z)| \leq g|z| + h$ for any $z \in \mathbb{R}$

as $n, p \to \infty$ with $p/n \to \gamma \in (0, \infty)$ for any compact set $\Lambda \subseteq (\lambda_{\min}, \infty)$,

$$\sup_{\lambda \in \Lambda} \left| \widehat{T}_\lambda^{\text{gcv}} - T_\lambda \right| \xrightarrow{\text{a.s.}} 0, \quad \text{and} \quad \sup_{\lambda \in \Lambda} \left| \widehat{T}_\lambda^{\text{loo}} - T_\lambda \right| \xrightarrow{\text{a.s.}} 0.$$

Remarks:

- Special case of $t(r) = r^2$ exploits bias-variance decomposition
- No bias-variance decomposition for general error functions and result requires a different proof technique via leave-one-out arguments
- Using uniformity arguments, the result can be extended for non-linear variational functionals

Linear functional estimation (uniform in λ)

Theorem. For error functions $t : \mathbb{R} \to \mathbb{R}$

1. that are differentiable
2. have derivative with linear growth rate, i.e., there exist constants $g, h > 0$ such that $|t'(z)| \leq g|z| + h$ for any $z \in \mathbb{R}$

as $n, p \to \infty$ with $p/n \to \gamma \in (0, \infty)$ for any compact set $\Lambda \subseteq (\lambda_{\min}, \infty)$,

$$\sup_{\lambda \in \Lambda} \left| \widehat{T}_\lambda^{\mathrm{gcv}} - T_\lambda \right| \xrightarrow{\text{a.s.}} 0, \quad \text{and} \quad \sup_{\lambda \in \Lambda} \left| \widehat{T}_\lambda^{\mathrm{loo}} - T_\lambda \right| \xrightarrow{\text{a.s.}} 0.$$

Remarks:

- Special case of $t(r) = r^2$ exploits bias-variance decomposition
- No bias-variance decomposition for general error functions and result requires a different proof technique via leave-one-out arguments
- Using uniformity arguments, the result can be extended for non-linear variational functionals

Linear functional estimation (uniform in λ)

Theorem. For error functions $t : \mathbb{R} \to \mathbb{R}$

1. that are differentiable
2. have derivative with linear growth rate, i.e., there exist constants $g, h > 0$ such that $|t'(z)| \leq g|z| + h$ for any $z \in \mathbb{R}$

as $n, p \to \infty$ with $p/n \to \gamma \in (0, \infty)$ for any compact set $\Lambda \subseteq (\lambda_{\min}, \infty)$,

$$\sup_{\lambda \in \Lambda} \big| \widehat{T}_\lambda^{\text{gcv}} - T_\lambda \big| \xrightarrow{\text{a.s.}} 0, \quad \text{and} \quad \sup_{\lambda \in \Lambda} \big| \widehat{T}_\lambda^{\text{loo}} - T_\lambda \big| \xrightarrow{\text{a.s.}} 0.$$

Remarks:

- Special case of $t(r) = r^2$ exploits bias-variance decomposition
- No bias-variance decomposition for general error functions and result requires a different proof technique via leave-one-out arguments
- Using uniformity arguments, the result can be extended for non-linear variational functionals

Linear functional estimation (uniform in λ)

Theorem. For error functions $t : \mathbb{R} \to \mathbb{R}$

1. that are differentiable
2. have derivative with linear growth rate, i.e., there exist constants $g, h > 0$ such that $|t'(z)| \leq g|z| + h$ for any $z \in \mathbb{R}$

as $n, p \to \infty$ with $p/n \to \gamma \in (0, \infty)$ for any compact set $\Lambda \subseteq (\lambda_{\mathrm{min}}, \infty)$,

$$\sup_{\lambda \in \Lambda} \left| \widehat{T}_{\lambda}^{\mathrm{gcv}} - T_{\lambda} \right| \xrightarrow{\mathrm{a.s.}} 0, \quad \text{and} \quad \sup_{\lambda \in \Lambda} \left| \widehat{T}_{\lambda}^{\mathrm{loo}} - T_{\lambda} \right| \xrightarrow{\mathrm{a.s.}} 0.$$

Remarks:

- Special case of $t(r) = r^2$ exploits bias-variance decomposition
- No bias-variance decomposition for general error functions and result requires a different proof technique via leave-one-out arguments
- Using uniformity arguments, the result can be extended for non-linear variational functionals

Linear functional estimation (uniform in λ)

Theorem. For error functions $t : \mathbb{R} \to \mathbb{R}$

1. that are differentiable
2. have derivative with linear growth rate, i.e., there exist constants $g, h > 0$ such that $|t'(z)| \leq g|z| + h$ for any $z \in \mathbb{R}$

as $n, p \to \infty$ with $p/n \to \gamma \in (0, \infty)$ for any compact set $\Lambda \subseteq (\lambda_{\min}, \infty)$,

$$\sup_{\lambda \in \Lambda} \left| \widehat{T}_\lambda^{\mathrm{gcv}} - T_\lambda \right| \xrightarrow{\text{a.s.}} 0, \quad \text{and} \quad \sup_{\lambda \in \Lambda} \left| \widehat{T}_\lambda^{\mathrm{loo}} - T_\lambda \right| \xrightarrow{\text{a.s.}} 0.$$

Remarks:

- Special case of $t(r) = r^2$ exploits bias-variance decomposition
- No bias-variance decomposition for general error functions and result requires a different proof technique via leave-one-out arguments
- Using uniformity arguments, the result can be extended for non-linear variational functionals

Application: quantile estimation

- Quantile of the out-of-sample error distribution:

$$Q_\lambda(\tau) = \mathrm{Quantile}(y_0 - x_0^T \widehat{\beta}_\lambda; \tau) = \underset{u \in \mathbb{R}}{\arg\min}\, \mathbb{E}\big[t_u(y_0 - x_0^T \widehat{\beta}_\lambda; \tau) \mid X, y\big]$$

  where $t_u(y - x_0^T \widehat{\beta}_\lambda; \tau)$ is τ-tiled pin-ball loss function with shift u
- Empirical quantiles $\widehat{Q}^{\mathrm{gcv}}$ and $\widehat{Q}^{\mathrm{loo}}$ (of $\widehat{P}_\lambda^{\mathrm{gcv}}$ and $\widehat{P}_\lambda^{\mathrm{loo}}$) $\xrightarrow{\text{a.s.}} Q_\lambda$
- Estimated quantiles can be used to construct prediction intervals:

$$\mathcal{I}_\lambda^{\mathrm{gcv}} = \big[x_0^\top \widehat{\beta}_\lambda - \widehat{Q}_\lambda^{\mathrm{gcv}}(\tau_l),\, x_0^\top \widehat{\beta}_\lambda + \widehat{Q}_\lambda^{\mathrm{gcv}}(\tau_u)\big] \quad \text{and} \quad \mathcal{I}_\lambda^{\mathrm{loo}}$$

Such intervals have correct coverage conditional on the training data:

Corollary. Under proportional asymptotics, almost surely

$$\mathbb{P}(y_0 \in \mathcal{I}_\lambda^{\mathrm{gcv}} \mid X, y) \xrightarrow{\text{a.s.}} 1 - \alpha, \quad \text{and} \quad \mathbb{P}(y_0 \in \mathcal{I}_\lambda^{\mathrm{loo}} \mid X, y) \xrightarrow{\text{a.s.}} 1 - \alpha.$$

Application: quantile estimation

- Quantile of the out-of-sample error distribution:

$$Q_\lambda(\tau) = \mathrm{Quantile}(y_0 - x_0^T \widehat{\beta}_\lambda; \tau) = \arg\min_{u \in \mathbb{R}} \mathbb{E}\big[t_u(y_0 - x_0^T \widehat{\beta}_\lambda; \tau) \mid X, y\big]$$

where $t_u(y - x_0^T \widehat{\beta}_\lambda; \tau)$ is τ-tiled pin-ball loss function with shift u

- Empirical quantiles $\widehat{Q}^{\mathrm{gcv}}$ and $\widehat{Q}^{\mathrm{loo}}$ (of $\widehat{P}_\lambda^{\mathrm{gcv}}$ and $\widehat{P}_\lambda^{\mathrm{loo}}$) $\xrightarrow{\text{a.s.}} Q_\lambda$
- Estimated quantiles can be used to construct prediction intervals:

$$\mathcal{I}_\lambda^{\mathrm{gcv}} = \big[x_0^\top \widehat{\beta}_\lambda - \widehat{Q}_\lambda^{\mathrm{gcv}}(\tau_l),\, x_0^\top \widehat{\beta}_\lambda + \widehat{Q}_\lambda^{\mathrm{gcv}}(\tau_u)\big] \quad \text{and} \quad \mathcal{I}_\lambda^{\mathrm{loo}}$$

Such intervals have correct coverage conditional on the training data:

Corollary. Under proportional asymptotics, almost surely

$$\mathbb{P}(y_0 \in \mathcal{I}_\lambda^{\mathrm{gcv}} \mid X, y) \xrightarrow{\text{a.s.}} 1 - \alpha, \quad \text{and} \quad \mathbb{P}(y_0 \in \mathcal{I}_\lambda^{\mathrm{loo}} \mid X, y) \xrightarrow{\text{a.s.}} 1 - \alpha.$$

Application: quantile estimation

- Quantile of the out-of-sample error distribution:

$$Q_\lambda(\tau) = \mathrm{Quantile}(y_0 - x_0^T \widehat{\beta}_\lambda; \tau) = \arg\min_{u \in \mathbb{R}} \mathbb{E}\big[t_u(y_0 - x_0^T \widehat{\beta}_\lambda; \tau) \mid X, y\big]$$

where $t_u(y - x_0^T \widehat{\beta}_\lambda; \tau)$ is τ-tiled pin-ball loss function with shift u

- Empirical quantiles $\widehat{Q}^{\mathrm{gcv}}$ and $\widehat{Q}^{\mathrm{loo}}$ (of $\widehat{P}_\lambda^{\mathrm{gcv}}$ and $\widehat{P}_\lambda^{\mathrm{loo}}$) $\xrightarrow{\text{a.s.}}$ Q_λ
- Estimated quantiles can be used to construct prediction intervals:

$$\mathcal{I}_\lambda^{\mathrm{gcv}} = \big[x_0^\top \widehat{\beta}_\lambda - \widehat{Q}_\lambda^{\mathrm{gcv}}(\tau_l),\ x_0^\top \widehat{\beta}_\lambda + \widehat{Q}_\lambda^{\mathrm{gcv}}(\tau_u)\big] \quad \text{and} \quad \mathcal{I}_\lambda^{\mathrm{loo}}$$

Such intervals have correct coverage conditional on the training data:

Corollary. Under proportional asymptotics, almost surely

$$\mathbb{P}(y_0 \in \mathcal{I}_\lambda^{\mathrm{gcv}} \mid X, y) \xrightarrow{\text{a.s.}} 1 - \alpha, \quad \text{and} \quad \mathbb{P}(y_0 \in \mathcal{I}_\lambda^{\mathrm{loo}} \mid X, y) \xrightarrow{\text{a.s.}} 1 - \alpha.$$

Application: quantile estimation

- Quantile of the out-of-sample error distribution:

$$Q_\lambda(\tau) = \mathrm{Quantile}(y_0 - x_0^T \widehat{\beta}_\lambda; \tau) = \arg\min_{u \in \mathbb{R}} \mathbb{E}\big[t_u(y_0 - x_0^T \widehat{\beta}_\lambda; \tau) \mid X, y\big]$$

  where $t_u(y - x_0^T \widehat{\beta}_\lambda; \tau)$ is τ-tiled pin-ball loss function with shift u

- Empirical quantiles $\widehat{Q}^{\mathrm{gcv}}$ and $\widehat{Q}^{\mathrm{loo}}$ (of $\widehat{P}_\lambda^{\mathrm{gcv}}$ and $\widehat{P}_\lambda^{\mathrm{loo}}$) $\xrightarrow{\text{a.s.}} Q_\lambda$
- Estimated quantiles can be used to construct prediction intervals:

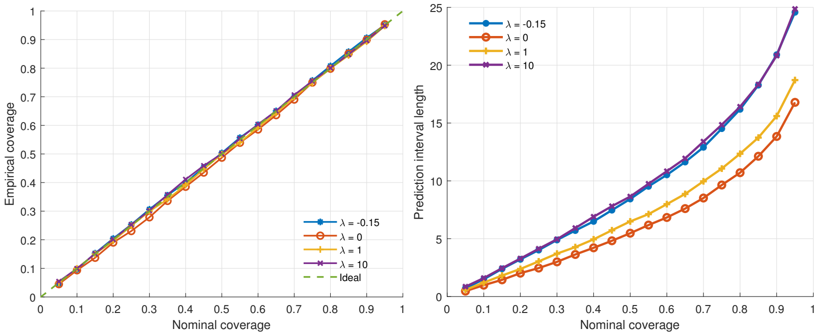
$$\mathcal{I}_\lambda^{\mathrm{gcv}} = \big[x_0^\top \widehat{\beta}_\lambda - \widehat{Q}_\lambda^{\mathrm{gcv}}(\tau_l),\, x_0^\top \widehat{\beta}_\lambda + \widehat{Q}_\lambda^{\mathrm{gcv}}(\tau_u)\big] \quad \text{and} \quad \mathcal{I}_\lambda^{\mathrm{loo}}$$

Such intervals have correct coverage conditional on the training data:

Corollary. Under proportional asymptotics, almost surely

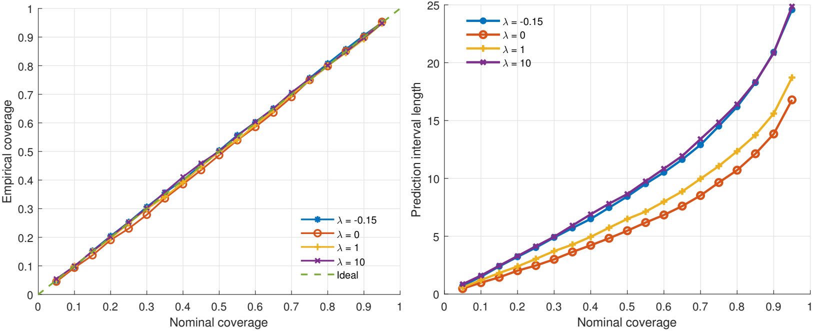
$$\mathbb{P}(y_0 \in \mathcal{I}_\lambda^{\mathrm{gcv}} \mid X, y) \xrightarrow{\text{a.s.}} 1 - \alpha, \quad \text{and} \quad \mathbb{P}(y_0 \in \mathcal{I}_\lambda^{\mathrm{loo}} \mid X, y) \xrightarrow{\text{a.s.}} 1 - \alpha.$$

Prediction intervals: illustration (coverage and length)



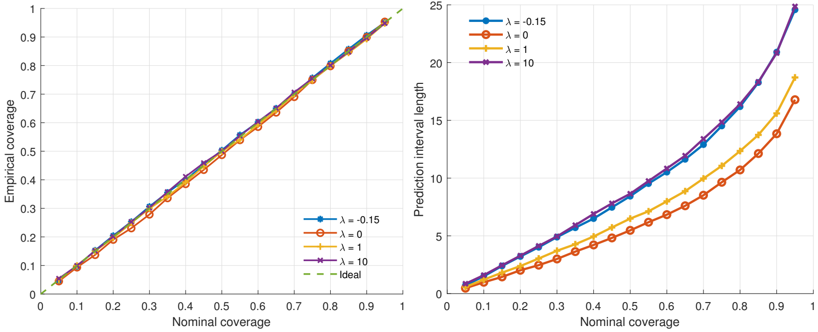
- $n = 2500$, $p = 5000$
- Features: autoregressive feature covariance structure
- Signal: latent signal aligned with the principal eigenvector
- Coverage nearly exact, even for $\lambda = 0$!
- The case of $\lambda = 0$ provides the minimum interval length!

Prediction intervals: illustration (coverage and length)



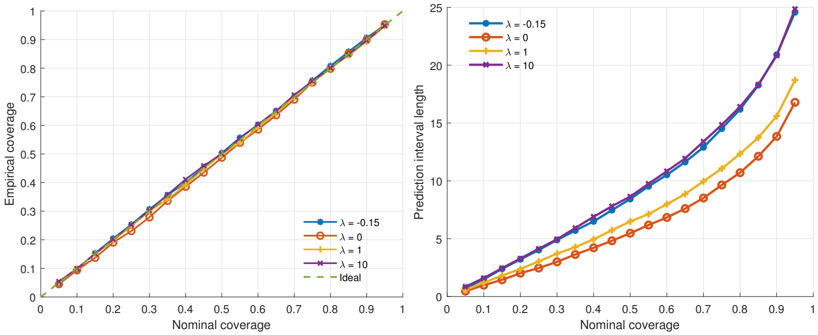
- $n = 2500$, $p = 5000$
- Features: autoregressive feature covariance structure
- Signal: latent signal aligned with the principal eigenvector
- Coverage nearly exact, even for $\lambda = 0$!
- The case of $\lambda = 0$ provides the minimum interval length!

Prediction intervals: illustration (coverage and length)



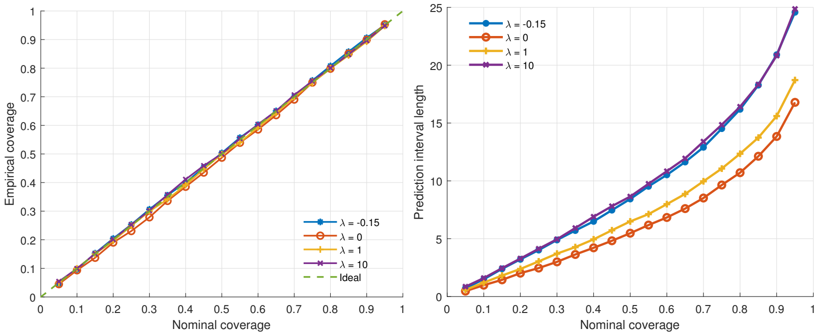
- $n = 2500$, $p = 5000$
- Features: autoregressive feature covariance structure
- Signal: latent signal aligned with the principal eigenvector
- Coverage nearly exact, even for $\lambda = 0$!
- The case of $\lambda = 0$ provides the minimum interval length!

Prediction intervals: illustration (coverage and length)



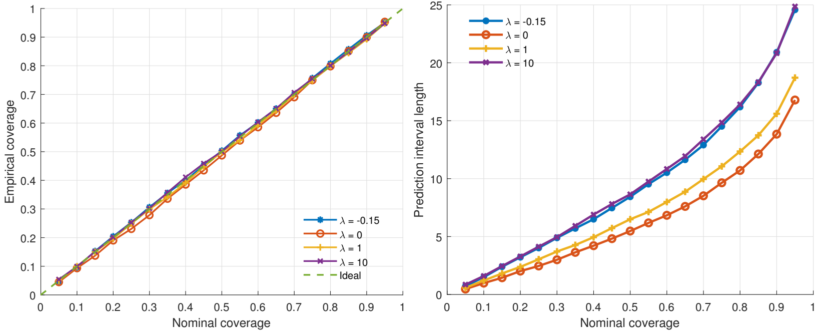
- $n = 2500$, $p = 5000$
- Features: autoregressive feature covariance structure
- Signal: latent signal aligned with the principal eigenvector
- Coverage nearly exact, even for $\lambda = 0$!
- The case of $\lambda = 0$ provides the minimum interval length!

Prediction intervals: illustration (coverage and length)



- $n = 2500$, $p = 5000$
- Features: autoregressive feature covariance structure
- Signal: latent signal aligned with the principal eigenvector
- Coverage nearly exact, even for $\lambda = 0$!
- The case of $\lambda = 0$ provides the minimum interval length!

Prediction intervals: illustration (coverage and length)



- $n = 2500$, $p = 5000$
- Features: autoregressive feature covariance structure
- Signal: latent signal aligned with the principal eigenvector
- Coverage nearly exact, even for $\lambda = 0$!
- The case of $\lambda = 0$ provides the minimum interval length!

Discussion and extensions

Take-away from this work: empirical distributions of GCV and LOOCV track out-of-sample error distribution and a wide class of its functionals for ridge regression under proportional asymptotics framework

Key relation that we exploit:

$$y_i - x_i^\top \widehat{\beta}_{-i,\lambda} = \frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - [L_\lambda]_{ii}} \approx \frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - \mathrm{tr}[L_\lambda]/n}$$

$$y_i - x_i^\top \widehat{\beta}_{-i,0} = \frac{[(XX^\top)^\dagger y]_i}{[(XX^\top)^\dagger]_{ii}} \approx \frac{[(XX^\top)^\dagger y]_i}{\mathrm{tr}[(XX^\top)^\dagger]/n}$$

Extensions:

- Generalized ridge/less regression through structural equivalences
- Kernel ridge/less regression through risk equivalences

Discussion and extensions

Take-away from this work: empirical distributions of GCV and LOOCV track out-of-sample error distribution and a wide class of its functionals for ridge regression under proportional asymptotics framework

Key relation that we exploit:

$$y_i - x_i^\top \widehat{\beta}_{-i,\lambda} = \frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - [L_\lambda]_{ii}} \approx \frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - \mathrm{tr}[L_\lambda]/n}$$

$$y_i - x_i^\top \widehat{\beta}_{-i,0} = \frac{[(XX^\top)^\dagger y]_i}{[(XX^\top)^\dagger]_{ii}} \approx \frac{[(XX^\top)^\dagger y]_i}{\mathrm{tr}[(XX^\top)^\dagger]/n}$$

Extensions:

- Generalized ridge/less regression through structural equivalences
- Kernel ridge/less regression through risk equivalences

Discussion and extensions

Take-away from this work: empirical distributions of GCV and LOOCV track out-of-sample error distribution and a wide class of its functionals for ridge regression under proportional asymptotics framework

Key relation that we exploit:

$$y_i - {x_i}^\top \widehat{\beta}_{-i,\lambda} = \frac{y_i - {x_i}^\top \widehat{\beta}_\lambda}{1 - [L_\lambda]_{ii}} \approx \frac{y_i - {x_i}^\top \widehat{\beta}_\lambda}{1 - \mathrm{tr}[L_\lambda]/n}$$

$$y_i - {x_i}^\top \widehat{\beta}_{-i,0} = \frac{[(XX^\top)^\dagger y]_i}{[(XX^\top)^\dagger]_{ii}} \approx \frac{[(XX^\top)^\dagger y]_i}{\mathrm{tr}[(XX^\top)^\dagger]/n}$$

Extensions:

- Generalized ridge/less regression through structural equivalences
- Kernel ridge/less regression through risk equivalences

Outline

Motivation and main punchlines

- When the data comprises of i.i.d. observations, we expect that more data will help in prediction or estimation.
- Double or multiple descent behaviour implies that for fixed feature size p (large value), as sample size increases the risk first decreases and then increases. More data can hurt!
- A procedure leading to worse risk as the number of observations increases is not using the data properly.

Key question: Can we modify any prediction procedure to mitigate the double or multiple descent behavior and achieve a monotonic risk behavior?

We propose two methods, dubbed zero-step and one-step, that take an input an arbitrary procedure and return a modified procedure that has a monotonic risk behavior. The main idea is that of subsampling.

Motivation and main punchlines

- When the data comprises of i.i.d. observations, we expect that more data will help in prediction or estimation.
- Double or multiple descent behaviour implies that for fixed feature size p (large value), as sample size increases the risk first decreases and then increases. More data can hurt!
- A procedure leading to worse risk as the number of observations increases is not using the data properly.

Key question: Can we modify any prediction procedure to mitigate the double or multiple descent behavior and achieve a monotonic risk behavior?

We propose two methods, dubbed zero-step and one-step, that take an input an arbitrary procedure and return a modified procedure that has a monotonic risk behavior. The main idea is that of subsampling.

Motivation and main punchlines

- When the data comprises of i.i.d. observations, we expect that more data will help in prediction or estimation.
- Double or multiple descent behaviour implies that for fixed feature size p (large value), as sample size increases the risk first decreases and then increases. More data can hurt!
- A procedure leading to worse risk as the number of observations increases is not using the data properly.

Key question: Can we modify any prediction procedure to mitigate the double or multiple descent behavior and achieve a monotonic risk behavior?

We propose two methods, dubbed zero-step and one-step, that take an input an arbitrary procedure and return a modified procedure that has a monotonic risk behavior. The main idea is that of subsampling.

Motivation and main punchlines

- When the data comprises of i.i.d. observations, we expect that more data will help in prediction or estimation.
- Double or multiple descent behaviour implies that for fixed feature size p (large value), as sample size increases the risk first decreases and then increases. More data can hurt!
- A procedure leading to worse risk as the number of observations increases is not using the data properly.

Key question: Can we modify any prediction procedure to mitigate the double or multiple descent behavior and achieve a monotonic risk behavior?

We propose two methods, dubbed zero-step and one-step, that take an input an arbitrary procedure and return a modified procedure that has a monotonic risk behavior. The main idea is that of subsampling.

Motivation and main punchlines

- When the data comprises of i.i.d. observations, we expect that more data will help in prediction or estimation.
- Double or multiple descent behaviour implies that for fixed feature size p (large value), as sample size increases the risk first decreases and then increases. More data can hurt!
- A procedure leading to worse risk as the number of observations increases is not using the data properly.

Key question: Can we modify any prediction procedure to mitigate the double or multiple descent behavior and achieve a monotonic risk behavior?

We propose two methods, dubbed zero-step and one-step, that take an input an arbitrary procedure and return a modified procedure that has a monotonic risk behavior. The main idea is that of subsampling.

Motivation and the problem

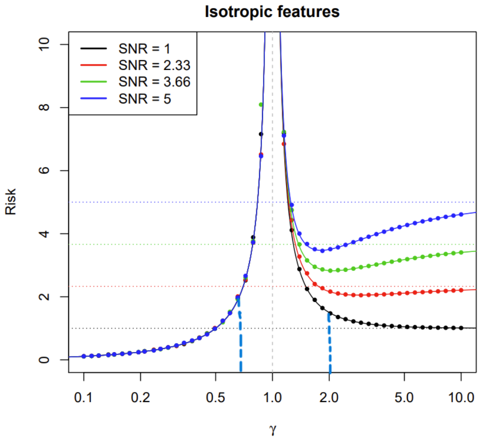


Figure: Risk of the minimum ℓ_2-norm least squares as a function of $p/n \approx \gamma$.

Hastie, Montanari, Rosset, Tibshirani, 2019: "Surprises in high-dimensional ridgeless least squares interpolation"

The problem

- Given a number of observations (n) and a number of features (p), how do we know if a lesser number of observations would actually yield a better risk?
- What is the best sample size to reduce the dataset in order to attain the best possible risk?

Solution: cross-validation.

The problem

- Given a number of observations (n) and a number of features (p), how do we know if a lesser number of observations would actually yield a better risk?
- What is the best sample size to reduce the dataset in order to attain the best possible risk?

Solution: cross-validation.

The problem

- Given a number of observations (n) and a number of features (p), how do we know if a lesser number of observations would actually yield a better risk?
- What is the best sample size to reduce the dataset in order to attain the best possible risk?

Solution: cross-validation.

Basic idea of zero-step procedure

Given any arbitrary prediction procedure at a given aspect ratio $\gamma = p/n$:

1. Risk estimation: construct a (dense grid of) aspect ratios $\geq \gamma$ by using datasets of sizes smaller than n, and estimate risks on test set
2. Model selection: select aspect ratio that delivers the smallest estimated risk and return the corresponding predictor
3. Risk monotonization: show that the risk profile of the resulting procedure is asymptotically monotone in the aspect ratio

Method highlights:

- applicable to generic (e.g black-box) prediction methods and common classification and regression loss functions
- model agnostic and requires minimal distributional assumptions
- works for procedures with diverging risks at some aspect ratios

Basic idea of zero-step procedure

Given any arbitrary prediction procedure at a given aspect ratio $\gamma = p/n$:

1. Risk estimation: construct a (dense grid of) aspect ratios $\geq \gamma$ by using datasets of sizes smaller than n, and estimate risks on test set
2. Model selection: select aspect ratio that delivers the smallest estimated risk and return the corresponding predictor
3. Risk monotonization: show that the risk profile of the resulting procedure is asymptotically monotone in the aspect ratio

Method highlights:

- applicable to generic (e.g black-box) prediction methods and common classification and regression loss functions
- model agnostic and requires minimal distributional assumptions
- works for procedures with diverging risks at some aspect ratios

Basic idea of zero-step procedure

Given any arbitrary prediction procedure at a given aspect ratio $\gamma = p/n$:

1. Risk estimation: construct a (dense grid of) aspect ratios $\geq \gamma$ by using datasets of sizes smaller than n, and estimate risks on test set
2. Model selection: select aspect ratio that delivers the smallest estimated risk and return the corresponding predictor
3. Risk monotonization: show that the risk profile of the resulting procedure is asymptotically monotone in the aspect ratio

Method highlights:

- applicable to generic (e.g black-box) prediction methods and common classification and regression loss functions
- model agnostic and requires minimal distributional assumptions
- works for procedures with diverging risks at some aspect ratios

Basic idea of zero-step procedure

Given any arbitrary prediction procedure at a given aspect ratio $\gamma = p/n$:

1. Risk estimation: construct a (dense grid of) aspect ratios $\geq \gamma$ by using datasets of sizes smaller than n, and estimate risks on test set
2. Model selection: select aspect ratio that delivers the smallest estimated risk and return the corresponding predictor
3. Risk monotonization: show that the risk profile of the resulting procedure is asymptotically monotone in the aspect ratio

Method highlights:

- applicable to generic (e.g black-box) prediction methods and common classification and regression loss functions
- model agnostic and requires minimal distributional assumptions
- works for procedures with diverging risks at some aspect ratios

Basic idea of zero-step procedure

Given any arbitrary prediction procedure at a given aspect ratio $\gamma = p/n$:

1. Risk estimation: construct a (dense grid of) aspect ratios $\geq \gamma$ by using datasets of sizes smaller than n, and estimate risks on test set
2. Model selection: select aspect ratio that delivers the smallest estimated risk and return the corresponding predictor
3. Risk monotonization: show that the risk profile of the resulting procedure is asymptotically monotone in the aspect ratio

Method highlights:

- applicable to generic (e.g black-box) prediction methods and common classification and regression loss functions
- model agnostic and requires minimal distributional assumptions
- works for procedures with diverging risks at some aspect ratios

Basic idea of zero-step procedure

Given any arbitrary prediction procedure at a given aspect ratio $\gamma = p/n$:

1. Risk estimation: construct a (dense grid of) aspect ratios $\geq \gamma$ by using datasets of sizes smaller than n, and estimate risks on test set
2. Model selection: select aspect ratio that delivers the smallest estimated risk and return the corresponding predictor
3. Risk monotonization: show that the risk profile of the resulting procedure is asymptotically monotone in the aspect ratio

Method highlights:

- applicable to generic (e.g black-box) prediction methods and common classification and regression loss functions
- model agnostic and requires minimal distributional assumptions
- works for procedures with diverging risks at some aspect ratios

Basic idea of zero-step procedure

Given any arbitrary prediction procedure at a given aspect ratio $\gamma = p/n$:

1. Risk estimation: construct a (dense grid of) aspect ratios $\geq \gamma$ by using datasets of sizes smaller than n, and estimate risks on test set
2. Model selection: select aspect ratio that delivers the smallest estimated risk and return the corresponding predictor
3. Risk monotonization: show that the risk profile of the resulting procedure is asymptotically monotone in the aspect ratio

Method highlights:

- applicable to generic (e.g black-box) prediction methods and common classification and regression loss functions
- model agnostic and requires minimal distributional assumptions
- works for procedures with diverging risks at some aspect ratios

Basic idea of zero-step procedure

Given any arbitrary prediction procedure at a given aspect ratio $\gamma = p/n$:

1. Risk estimation: construct a (dense grid of) aspect ratios $\geq \gamma$ by using datasets of sizes smaller than n, and estimate risks on test set
2. Model selection: select aspect ratio that delivers the smallest estimated risk and return the corresponding predictor
3. Risk monotonization: show that the risk profile of the resulting procedure is asymptotically monotone in the aspect ratio

Method highlights:

- applicable to generic (e.g black-box) prediction methods and common classification and regression loss functions
- model agnostic and requires minimal distributional assumptions
- works for procedures with diverging risks at some aspect ratios

Basic idea of zero-step procedure

Given any arbitrary prediction procedure at a given aspect ratio $\gamma = p/n$:

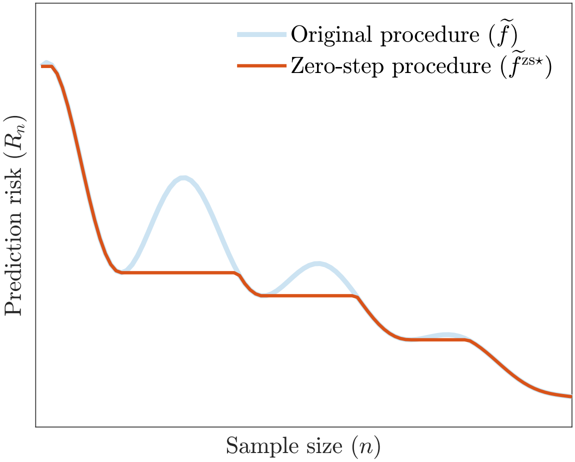
1. Risk estimation: construct a (dense grid of) aspect ratios $\geq \gamma$ by using datasets of sizes smaller than n, and estimate risks on test set
2. Model selection: select aspect ratio that delivers the smallest estimated risk and return the corresponding predictor
3. Risk monotonization: show that the risk profile of the resulting procedure is asymptotically monotone in the aspect ratio

Method highlights:

- applicable to generic (e.g black-box) prediction methods and common classification and regression loss functions
- model agnostic and requires minimal distributional assumptions
- works for procedures with diverging risks at some aspect ratios

Risk monotonization illustration

If R_n represents the "risk" of a procedure at sample size n, then by risk monotonization we mean a procedure with risk $\min_{m \leq n} R_m$.



Risk monotonization guarantee

Theorem. Under the proportional asymptotics regime ($p/n \to \gamma$), and a mild assumption on the convergence of the prediction risk of $\widehat{f}$ trained on datasets with a limiting aspect ratio ζ converges to $R^{\mathrm{det}}(\zeta; \widehat{f})$, we show:

$$R(\widehat{f}^{\mathrm{cv}}) \; = \; \inf_{\zeta \in [\gamma, \infty]} R^{\mathrm{det}}(\zeta; \widehat{f}) \; \times \; (1 + o_p(1)).$$

This shows that the zero-step predictor has a monotone risk in terms of the sample size and hence with respect to the limiting aspect ratio.

This is a model-free result in that no parametric model is assumed for the data. This is unlike most results in overparametrized learning which require stringent assumptions.

Risk monotonization guarantee

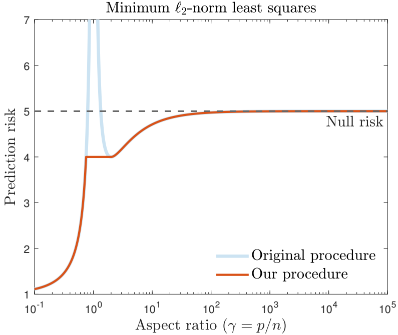
Theorem. Under the proportional asymptotics regime ($p/n \to \gamma$), and a mild assumption on the convergence of the prediction risk of $\widehat{f}$ trained on datasets with a limiting aspect ratio ζ converges to $R^{\text{det}}(\zeta; \widehat{f})$, we show:

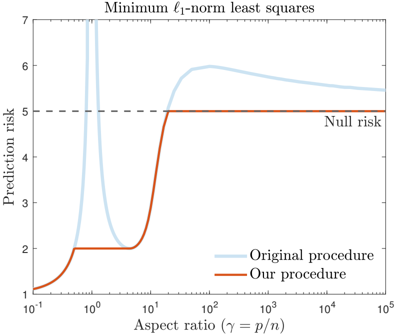
$$R(\widehat{f}^{\text{cv}}) \;=\; \inf_{\zeta \in [\gamma, \infty]} R^{\text{det}}(\zeta; \widehat{f}) \;\times\; (1 + o_p(1)).$$

This shows that the zero-step predictor has a monotone risk in terms of the sample size and hence with respect to the limiting aspect ratio.

This is a model-free result in that no parametric model is assumed for the data. This is unlike most results in overparametrized learning which require stringent assumptions.

Risk monotonization (illustration)





- minimum ℓ_2-norm least squares (ridgeless regression)
- minimum ℓ_1-norm least squares (lassoless regression)

Discussion and extensions

Take-aways:

- We have introduced the zero-step prediction procedure that provably monotonizes the risk of a given predictor.
- The main idea is cross-validation based on test data, but with splitting done so as to maintain the limiting aspect ratio.

Extensions:

- We also introduce a one-step prediction procedure inspired by classical one-step estimator that improves on zero-step procedure (similar to boosting)
- Both zero-step and one-step procedures can be further improved by multiple subsamplings and averaging (similar to bagging)

Discussion and extensions

Take-aways:

- We have introduced the zero-step prediction procedure that provably monotonizes the risk of a given predictor.
- The main idea is cross-validation based on test data, but with splitting done so as to maintain the limiting aspect ratio.

Extensions:

- We also introduce a one-step prediction procedure inspired by classical one-step estimator that improves on zero-step procedure (similar to boosting)
- Both zero-step and one-step procedures can be further improved by multiple subsamplings and averaging (similar to bagging)

Discussion and extensions

Take-aways:

- We have introduced the zero-step prediction procedure that provably monotonizes the risk of a given predictor.
- The main idea is cross-validation based on test data, but with splitting done so as to maintain the limiting aspect ratio.

Extensions:

- We also introduce a one-step prediction procedure inspired by classical one-step estimator that improves on zero-step procedure (similar to boosting)
- Both zero-step and one-step procedures can be further improved by multiple subsamplings and averaging (similar to bagging)

Discussion and extensions

Take-aways:

- We have introduced the zero-step prediction procedure that provably monotonizes the risk of a given predictor.
- The main idea is cross-validation based on test data, but with splitting done so as to maintain the limiting aspect ratio.

Extensions:

- We also introduce a one-step prediction procedure inspired by classical one-step estimator that improves on zero-step procedure (similar to boosting)
- Both zero-step and one-step procedures can be further improved by multiple subsamplings and averaging (similar to bagging)

Discussion and extensions

Take-aways:

- We have introduced the zero-step prediction procedure that provably monotonizes the risk of a given predictor.
- The main idea is cross-validation based on test data, but with splitting done so as to maintain the limiting aspect ratio.

Extensions:

- We also introduce a one-step prediction procedure inspired by classical one-step estimator that improves on zero-step procedure (similar to boosting)
- Both zero-step and one-step procedures can be further improved by multiple subsamplings and averaging (similar to bagging)

Discussion and extensions

Take-aways:

- We have introduced the zero-step prediction procedure that provably monotonizes the risk of a given predictor.
- The main idea is cross-validation based on test data, but with splitting done so as to maintain the limiting aspect ratio.

Extensions:

- We also introduce a one-step prediction procedure inspired by classical one-step estimator that improves on zero-step procedure (similar to boosting)
- Both zero-step and one-step procedures can be further improved by multiple subsamplings and averaging (similar to bagging)

Outline

Motivation and main punchlines

Key question: is there a principled measure of model complexity in general for overparameterized models?

- Propose measures of model complexity that are:
  - algorithm-specific and applies for any prediction algorithm
  - produce a number between 0 and n (the number of observations)
- Two variants of model complexities are:
  - emergent model complexity that depends on the prediction algorithm as well as underlying the regression function
  - intrinsic model complexity that depends on the prediction algorithm only and its adaptability to pure noise

Based on ideas from optimism theory and degrees of freedom.

Motivation and main punchlines

Key question: is there a principled measure of model complexity in general for overparameterized models?

- Propose measures of model complexity that are:
  - algorithm-specific and applies for any prediction algorithm
  - produce a number between 0 and n (the number of observations)
- Two variants of model complexities are:
  - emergent model complexity that depends on the prediction algorithm as well as underlying the regression function
  - intrinsic model complexity that depends on the prediction algorithm only and its adaptability to pure noise

Based on ideas from optimism theory and degrees of freedom.

Motivation and main punchlines

Key question: is there a principled measure of model complexity in general for overparameterized models?

- Propose measures of model complexity that are:
  - algorithm-specific and applies for any prediction algorithm
  - produce a number between 0 and n (the number of observations)
- Two variants of model complexities are:
  - emergent model complexity that depends on the prediction algorithm as well as underlying the regression function
  - intrinsic model complexity that depends on the prediction algorithm only and its adaptability to pure noise

Based on ideas from optimism theory and degrees of freedom.

Motivation and main punchlines

Key question: is there a principled measure of model complexity in general for overparameterized models?

- Propose measures of model complexity that are:
  - algorithm-specific and applies for any prediction algorithm
  - produce a number between 0 and n (the number of observations)
- Two variants of model complexities are:
  - emergent model complexity that depends on the prediction algorithm as well as underlying the regression function
  - intrinsic model complexity that depends on the prediction algorithm only and its adaptability to pure noise

Based on ideas from optimism theory and degrees of freedom.

Motivation and main punchlines

Key question: is there a principled measure of model complexity in general for overparameterized models?

- Propose measures of model complexity that are:
  - algorithm-specific and applies for any prediction algorithm
  - produce a number between 0 and n (the number of observations)
- Two variants of model complexities are:
  - emergent model complexity that depends on the prediction algorithm as well as underlying the regression function
  - intrinsic model complexity that depends on the prediction algorithm only and its adaptability to pure noise

Based on ideas from optimism theory and degrees of freedom.

Motivation and main punchlines

Key question: is there a principled measure of model complexity in general for overparameterized models?

- Propose measures of model complexity that are:
  - algorithm-specific and applies for any prediction algorithm
  - produce a number between 0 and n (the number of observations)
- Two variants of model complexities are:
  - emergent model complexity that depends on the prediction algorithm as well as underlying the regression function
  - intrinsic model complexity that depends on the prediction algorithm only and its adaptability to pure noise

Based on ideas from optimism theory and degrees of freedom.

Motivation and main punchlines

Key question: is there a principled measure of model complexity in general for overparameterized models?

- Propose measures of model complexity that are:
  - algorithm-specific and applies for any prediction algorithm
  - produce a number between 0 and n (the number of observations)
- Two variants of model complexities are:
  - emergent model complexity that depends on the prediction algorithm as well as underlying the regression function
  - intrinsic model complexity that depends on the prediction algorithm only and its adaptability to pure noise

Based on ideas from optimism theory and degrees of freedom.

Fixed-X degrees of freedom

Consider data $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$, $i = 1, \ldots, n$ such that $y_i = f(x_i) + \varepsilon_i$ where $f : \mathbb{R}^p \to \mathbb{R}$ is regression function, ε_i has mean 0 and variance σ^2.

Let $\mathcal{A}$ be any fitting algorithm that maps $\{(x_i, y_i)\}_{i=1}^n \overset{\mathcal{A}}{\mapsto} \widehat{f}$.
The degrees of freedom of predictor $\widehat{f}$ is defined as

$$\mathrm{DofF}(\widehat{f}) = \sum_{i=1}^n \mathrm{Cov}(y_i, \widehat{f}(x_i))/\sigma^2 = \mathrm{tr}\left[\mathrm{Cov}(y, \widehat{f}(X))\right]/\sigma^2,$$

where y: response vector, X: feature matrix, $\widehat{f}(X)$: predicted response

Where does squared error loss come into play?

$$\underbrace{\mathbb{E}\Big[\sum_{i=1}^n (\widetilde{y}_i - \widehat{f}(x_i))^2\Big]}_{\text{fixed-X prediction error } =: \mathrm{ErrF}(\widehat{f})} - \underbrace{\mathbb{E}\Big[\sum_{i=1}^n (y_i - \widehat{f}(x_i))^2\Big]}_{\text{expected training error } =: \mathrm{ErrT}(\widehat{f})} = 2\sigma^2 \mathrm{DofF}(\widehat{f})$$

Fixed-X degrees of freedom

Consider data $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$, $i = 1, \ldots, n$ such that $y_i = f(x_i) + \varepsilon_i$ where $f : \mathbb{R}^p \to \mathbb{R}$ is regression function, ε_i has mean 0 and variance σ^2.

Let $\mathcal{A}$ be any fitting algorithm that maps $\{(x_i, y_i)\}_{i=1}^n \overset{\mathcal{A}}{\mapsto} \widehat{f}$.
The degrees of freedom of predictor $\widehat{f}$ is defined as

$$\mathrm{DofF}(\widehat{f}) = \sum_{i=1}^{n} \mathrm{Cov}(y_i, \widehat{f}(x_i))/\sigma^2 = \mathrm{tr}\left[\mathrm{Cov}(y, \widehat{f}(X))\right]/\sigma^2,$$

where y: response vector, X: feature matrix, $\widehat{f}(X)$: predicted response

Where does squared error loss come into play?

$$\underbrace{\mathbb{E}\Big[\sum_{i=1}^{n}(\widetilde{y}_i - \widehat{f}(x_i))^2\Big]}_{\text{fixed-X prediction error } =: \mathrm{ErrF}(\widehat{f})} \quad - \quad \underbrace{\mathbb{E}\Big[\sum_{i=1}^{n}(y_i - \widehat{f}(x_i))^2\Big]}_{\text{expected training error } =: \mathrm{ErrT}(\widehat{f})} \quad = 2\sigma^2 \mathrm{DofF}(\widehat{f})$$

Fixed-X degrees of freedom

Consider data $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$, $i = 1, \dots, n$ such that $y_i = f(x_i) + \varepsilon_i$ where $f : \mathbb{R}^p \to \mathbb{R}$ is regression function, ε_i has mean 0 and variance σ^2.

Let $\mathcal{A}$ be any fitting algorithm that maps $\{(x_i, y_i)\}_{i=1}^n \overset{\mathcal{A}}{\mapsto} \widehat{f}$.
The degrees of freedom of predictor $\widehat{f}$ is defined as

$$\mathrm{DofF}(\widehat{f}) = \sum_{i=1}^{n} \mathrm{Cov}(y_i, \widehat{f}(x_i))/\sigma^2 = \mathrm{tr}\left[\,\mathrm{Cov}(y, \widehat{f}(X))\right]/\sigma^2,$$

where y: response vector, X: feature matrix, $\widehat{f}(X)$: predicted response

Where does squared error loss come into play?

$$\underbrace{\mathbb{E}\Big[\sum_{i=1}^{n}(\widetilde{y}_i - \widehat{f}(x_i))^2\Big]}_{\text{fixed-X prediction error } =: \mathrm{ErrF}(\widehat{f})} - \underbrace{\mathbb{E}\Big[\sum_{i=1}^{n}(y_i - \widehat{f}(x_i))^2\Big]}_{\text{expected training error } =: \mathrm{ErrT}(\widehat{f})} = 2\sigma^2 \mathrm{DofF}(\widehat{f})$$

Fixed-X degrees of freedom

Consider data $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$, $i = 1, \ldots, n$ such that $y_i = f(x_i) + \varepsilon_i$ where $f : \mathbb{R}^p \to \mathbb{R}$ is regression function, ε_i has mean 0 and variance σ^2.

Let $\mathcal{A}$ be any fitting algorithm that maps $\{(x_i, y_i)\}_{i=1}^n \overset{\mathcal{A}}{\mapsto} \widehat{f}$.
The degrees of freedom of predictor $\widehat{f}$ is defined as

$$\mathrm{DofF}(\widehat{f}) = \sum_{i=1}^{n} \mathrm{Cov}(y_i, \widehat{f}(x_i))/\sigma^2 = \mathrm{tr}\left[\mathrm{Cov}(y, \widehat{f}(X))\right]/\sigma^2,$$

where y: response vector, X: feature matrix, $\widehat{f}(X)$: predicted response

Where does squared error loss come into play?

$$\underbrace{\mathbb{E}\Big[\sum_{i=1}^{n}(\widetilde{y}_i - \widehat{f}(x_i))^2\Big]}_{\text{fixed-X prediction error } =:\mathrm{ErrF}(\widehat{f})} - \underbrace{\mathbb{E}\Big[\sum_{i=1}^{n}(y_i - \widehat{f}(x_i))^2\Big]}_{\text{expected training error } =:\mathrm{ErrT}(\widehat{f})} = 2\sigma^2\mathrm{DofF}(\widehat{f})$$

Fixed-X degrees of freedom

Consider data $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$, $i = 1, \ldots, n$ such that $y_i = f(x_i) + \varepsilon_i$ where $f : \mathbb{R}^p \to \mathbb{R}$ is regression function, ε_i has mean 0 and variance σ^2.

Let $\mathcal{A}$ be any fitting algorithm that maps $\{(x_i, y_i)\}_{i=1}^n \overset{\mathcal{A}}{\mapsto} \widehat{f}$.
The degrees of freedom of predictor $\widehat{f}$ is defined as

$$\mathrm{DofF}(\widehat{f}) = \sum_{i=1}^n \mathrm{Cov}(y_i, \widehat{f}(x_i))/\sigma^2 = \mathrm{tr}\left[\,\mathrm{Cov}(y, \widehat{f}(X))\right]/\sigma^2,$$

where y: response vector, X: feature matrix, $\widehat{f}(X)$: predicted response

Where does squared error loss come into play?

$$\underbrace{\mathbb{E}\Big[\sum_{i=1}^n (\widetilde{y}_i - \widehat{f}(x_i))^2\Big]}_{\text{fixed-X prediction error } =: \mathrm{ErrF}(\widehat{f})} \quad - \quad \underbrace{\mathbb{E}\Big[\sum_{i=1}^n (y_i - \widehat{f}(x_i))^2\Big]}_{\text{expected training error } =: \mathrm{ErrT}(\widehat{f})} \quad = 2\sigma^2 \mathrm{DofF}(\widehat{f})$$

Fixed-X degrees of freedom in linear regression

- Suppose $p \le n$ and X has full (column) rank, and we take $\widehat{f}$ to be ordinary least squares predictor $\widehat{f}(X) = X\widehat{\beta}$, where

$$\widehat{\beta} = \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2 = (X^T X)^{-1} X^T y.$$

$$\mathrm{DofF}(\widehat{f}) = \mathrm{tr}[\mathrm{Cov}(y, X\widehat{\beta})]/\sigma^2 = \mathrm{tr}[\sigma^2 X (X^T X)^{-1} X^T]/\sigma^2 = p.$$

- Suppose $p \ge n$ and X has full (row) rank, and we take $\widehat{f}$ to be min ℓ_2-norm least squares predictor $\widehat{f}(X) = X\widehat{\beta}$, where

$$\widehat{\beta} = \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p} \{\|\beta\|_2 : X\beta = y\} = (X^\top X)^\dagger X^\top y.$$

$$\mathrm{DofF}(\widehat{f}) = \mathrm{tr}[\mathrm{Cov}(y, X\widehat{\beta})]/\sigma^2 = \mathrm{tr}[\sigma^2 X^\top (X X^\top)^{-1} X]/\sigma^2 = n.$$

Thus, $\mathrm{DofF}(\widehat{f})$ is p for $p \le n$, but is always n for $p \ge n$ (not meaningful).

This is fixed-X degrees of freedom. How to extend for random-X setting?

Fixed-X degrees of freedom in linear regression

- Suppose $p \le n$ and X has full (column) rank, and we take $\widehat{f}$ to be ordinary least squares predictor $\widehat{f}(X) = X\widehat{\beta}$, where

$$\widehat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \|y - X\beta\|_2^2 = (X^T X)^{-1} X^T y.$$

$$\mathrm{DofF}(\widehat{f}) = \mathrm{tr}[\mathrm{Cov}(y, X\widehat{\beta})]/\sigma^2 = \mathrm{tr}[\sigma^2 X (X^T X)^{-1} X^T]/\sigma^2 = p.$$

- Suppose $p \ge n$ and X has full (row) rank, and we take $\widehat{f}$ to be min ℓ_2-norm least squares predictor $\widehat{f}(X) = X\widehat{\beta}$, where

$$\widehat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \{\|\beta\|_2 : X\beta = y\} = (X^\top X)^\dagger X^\top y.$$

$$\mathrm{DofF}(\widehat{f}) = \mathrm{tr}[\mathrm{Cov}(y, X\widehat{\beta})]/\sigma^2 = \mathrm{tr}[\sigma^2 X^\top (XX^\top)^{-1} X]/\sigma^2 = n.$$

Thus, $\mathrm{DofF}(\widehat{f})$ is p for $p \le n$, but is always n for $p \ge n$ (not meaningful).

This is fixed-X degrees of freedom. How to extend for random-X setting?

Fixed-X degrees of freedom in linear regression

- Suppose $p \leq n$ and X has full (column) rank, and we take $\widehat{f}$ to be ordinary least squares predictor $\widehat{f}(X) = X\widehat{\beta}$, where

$$\widehat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2 = (X^T X)^{-1} X^T y.$$

$$\mathrm{DofF}(\widehat{f}) = \mathrm{tr}[\mathrm{Cov}(y, X\widehat{\beta})]/\sigma^2 = \mathrm{tr}[\sigma^2 X(X^T X)^{-1} X^T]/\sigma^2 = p.$$

- Suppose $p \geq n$ and X has full (row) rank, and we take $\widehat{f}$ to be min ℓ_2-norm least squares predictor $\widehat{f}(X) = X\widehat{\beta}$, where

$$\widehat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \{\|\beta\|_2 : X\beta = y\} = (X^\top X)^\dagger X^\top y.$$

$$\mathrm{DofF}(\widehat{f}) = \mathrm{tr}[\mathrm{Cov}(y, X\widehat{\beta})]/\sigma^2 = \mathrm{tr}[\sigma^2 X^\top (XX^\top)^{-1} X]/\sigma^2 = n.$$

Thus, $\mathrm{DofF}(\widehat{f})$ is p for $p \leq n$, but is always n for $p \geq n$ (not meaningful).

This is fixed-X degrees of freedom. How to extend for random-X setting?

Fixed-X degrees of freedom in linear regression

- Suppose $p \le n$ and X has full (column) rank, and we take $\widehat{f}$ to be ordinary least squares predictor $\widehat{f}(X) = X\widehat{\beta}$, where

$$\widehat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \|y - X\beta\|_2^2 = (X^T X)^{-1} X^T y.$$

$$\mathrm{DofF}(\widehat{f}) = \mathrm{tr}[\mathrm{Cov}(y, X\widehat{\beta})]/\sigma^2 = \mathrm{tr}[\sigma^2 X (X^T X)^{-1} X^T]/\sigma^2 = p.$$

- Suppose $p \ge n$ and X has full (row) rank, and we take $\widehat{f}$ to be min ℓ_2-norm least squares predictor $\widehat{f}(X) = X\widehat{\beta}$, where

$$\widehat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \{\|\beta\|_2 : X\beta = y\} = (X^\top X)^\dagger X^\top y.$$

$$\mathrm{DofF}(\widehat{f}) = \mathrm{tr}[\mathrm{Cov}(y, X\widehat{\beta})]/\sigma^2 = \mathrm{tr}[\sigma^2 X^\top (X X^\top)^{-1} X]/\sigma^2 = n.$$

Thus, $\mathrm{DofF}(\widehat{f})$ is p for $p \le n$, but is always n for $p \ge n$ (not meaningful).

This is fixed-X degrees of freedom. How to extend for random-X setting?

Fixed-X degrees of freedom in linear regression

- Suppose $p \le n$ and X has full (column) rank, and we take $\widehat{f}$ to be ordinary least squares predictor $\widehat{f}(X) = X\widehat{\beta}$, where

$$\widehat{\beta} = \mathop{\arg\min}_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2 = (X^T X)^{-1} X^T y.$$

$$\mathrm{DofF}(\widehat{f}) = \mathrm{tr}[\mathrm{Cov}(y, X\widehat{\beta})]/\sigma^2 = \mathrm{tr}[\sigma^2 X (X^T X)^{-1} X^T]/\sigma^2 = p.$$

- Suppose $p \ge n$ and X has full (row) rank, and we take $\widehat{f}$ to be min ℓ_2-norm least squares predictor $\widehat{f}(X) = X\widehat{\beta}$, where

$$\widehat{\beta} = \mathop{\arg\min}_{\beta \in \mathbb{R}^p} \{\|\beta\|_2 : X\beta = y\} = (X^\top X)^\dagger X^\top y.$$

$$\mathrm{DofF}(\widehat{f}) = \mathrm{tr}[\mathrm{Cov}(y, X\widehat{\beta})]/\sigma^2 = \mathrm{tr}[\sigma^2 X^\top (X X^\top)^{-1} X]/\sigma^2 = n.$$

Thus, $\mathrm{DofF}(\widehat{f})$ is p for $p \le n$, but is always n for $p \ge n$ (not meaningful).

This is fixed-X degrees of freedom. How to extend for random-X setting?

Fixed-X degrees of freedom in linear regression

- Suppose $p \leq n$ and X has full (column) rank, and we take $\widehat{f}$ to be ordinary least squares predictor $\widehat{f}(X) = X\widehat{\beta}$, where

$$\widehat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2 = (X^T X)^{-1} X^T y.$$

$$\mathrm{DofF}(\widehat{f}) = \mathrm{tr}[\mathrm{Cov}(y, X\widehat{\beta})]/\sigma^2 = \mathrm{tr}[\sigma^2 X (X^T X)^{-1} X^T]/\sigma^2 = p.$$

- Suppose $p \geq n$ and X has full (row) rank, and we take $\widehat{f}$ to be min ℓ_2-norm least squares predictor $\widehat{f}(X) = X\widehat{\beta}$, where

$$\widehat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \{\|\beta\|_2 : X\beta = y\} = (X^\top X)^\dagger X^\top y.$$

$$\mathrm{DofF}(\widehat{f}) = \mathrm{tr}[\mathrm{Cov}(y, X\widehat{\beta})]/\sigma^2 = \mathrm{tr}[\sigma^2 X^\top (X X^\top)^{-1} X]/\sigma^2 = n.$$

Thus, $\mathrm{DofF}(\widehat{f})$ is p for $p \leq n$, but is always n for $p \geq n$ (not meaningful).

This is fixed-X degrees of freedom. How to extend for random-X setting?

Fixed-X degrees of freedom in linear regression

- Suppose $p \leq n$ and X has full (column) rank, and we take $\widehat{f}$ to be ordinary least squares predictor $\widehat{f}(X) = X\widehat{\beta}$, where

$$\widehat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2 = (X^T X)^{-1} X^T y.$$

$$\mathrm{DofF}(\widehat{f}) = \mathrm{tr}[\mathrm{Cov}(y, X\widehat{\beta})]/\sigma^2 = \mathrm{tr}[\sigma^2 X (X^T X)^{-1} X^T]/\sigma^2 = p.$$

- Suppose $p \geq n$ and X has full (row) rank, and we take $\widehat{f}$ to be min ℓ_2-norm least squares predictor $\widehat{f}(X) = X\widehat{\beta}$, where

$$\widehat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \{\|\beta\|_2 : X\beta = y\} = (X^\top X)^\dagger X^\top y.$$

$$\mathrm{DofF}(\widehat{f}) = \mathrm{tr}[\mathrm{Cov}(y, X\widehat{\beta})]/\sigma^2 = \mathrm{tr}[\sigma^2 X^\top (X X^\top)^{-1} X]/\sigma^2 = n.$$

Thus, $\mathrm{DofF}(\widehat{f})$ is p for $p \leq n$, but is always n for $p \geq n$ (not meaningful).

This is fixed-X degrees of freedom. How to extend for random-X setting?

Re-interpreting fixed-X degrees of freedom

Fixed-X degrees of freedom is a standard algorithm specific measure of complexity, but no notion of random-X degrees of freedom we know of.

We cast fixed-X degrees of freedom from a different perspective.

- Define fixed-X optimism of $\widehat{f}$ by $\mathrm{OptF}(\widehat{f}) = \mathrm{ErrF}(\widehat{f}) - \mathrm{ErrT}(\widehat{f})$.
- Consider the following family of "reference" models:
  - $\mathcal{A}^{\mathrm{ref}}$ is the least squares reference algorithm,
  - (U_k, v) is random design with k features, and noise with level σ^2.
- Recall that $\mathrm{DofF}(\mathcal{A}^{\mathrm{ref}}(U_k, v)) = k$ so long as $\mathrm{rank}(U_k) = k$.
- Thus, for a fitting procedure $\widehat{f} = \mathcal{A}(X, y)$, $\mathrm{DofF}(\widehat{f})$ is also equal to the value of k that satisfy the following relation:

$$\mathrm{OptF}(\mathcal{A}(X, y)) = \mathrm{OptF}(\mathcal{A}^{\mathrm{ref}}(U_k, v)) \qquad \text{(dfF)}$$

Re-interpreting fixed-X degrees of freedom

Fixed-X degrees of freedom is a standard algorithm specific measure of complexity, but no notion of random-X degrees of freedom we know of.

We cast fixed-X degrees of freedom from a different perspective.

- Define fixed-X optimism of $\widehat{f}$ by $\mathrm{OptF}(\widehat{f}) = \mathrm{ErrF}(\widehat{f}) - \mathrm{ErrT}(\widehat{f})$.
- Consider the following family of "reference" models:
  - $\mathcal{A}^{\mathrm{ref}}$ is the least squares reference algorithm,
  - (U_k, v) is random design with k features, and noise with level σ^2.
- Recall that $\mathrm{DofF}(\mathcal{A}^{\mathrm{ref}}(U_k, v)) = k$ so long as $\mathrm{rank}(U_k) = k$.
- Thus, for a fitting procedure $\widehat{f} = \mathcal{A}(X, y)$, $\mathrm{DofF}(\widehat{f})$ is also equal to the value of k that satisfy the following relation:

$$\mathrm{OptF}(\mathcal{A}(X, y)) = \mathrm{OptF}(\mathcal{A}^{\mathrm{ref}}(U_k, v)) \qquad \text{(dfF)}$$

Re-interpreting fixed-X degrees of freedom

Fixed-X degrees of freedom is a standard algorithm specific measure of complexity, but no notion of random-X degrees of freedom we know of.

We cast fixed-X degrees of freedom from a different perspective.

- Define fixed-X optimism of $\widehat{f}$ by $\mathrm{OptF}(\widehat{f}) = \mathrm{ErrF}(\widehat{f}) - \mathrm{ErrT}(\widehat{f})$.
- Consider the following family of "reference" models:
  - $\mathcal{A}^{\mathrm{ref}}$ is the least squares reference algorithm,
  - (U_k, v) is random design with k features, and noise with level σ^2.
- Recall that $\mathrm{DofF}(\mathcal{A}^{\mathrm{ref}}(U_k, v)) = k$ so long as $\mathrm{rank}(U_k) = k$.
- Thus, for a fitting procedure $\widehat{f} = \mathcal{A}(X, y)$, $\mathrm{DofF}(\widehat{f})$ is also equal to the value of k that satisfy the following relation:

$$\mathrm{OptF}(\mathcal{A}(X, y)) = \mathrm{OptF}(\mathcal{A}^{\mathrm{ref}}(U_k, v)) \qquad \text{(dfF)}$$

Re-interpreting fixed-X degrees of freedom

Fixed-X degrees of freedom is a standard algorithm specific measure of complexity, but no notion of random-X degrees of freedom we know of.

We cast fixed-X degrees of freedom from a different perspective.

- Define fixed-X optimism of $\widehat{f}$ by $\mathrm{OptF}(\widehat{f}) = \mathrm{ErrF}(\widehat{f}) - \mathrm{ErrT}(\widehat{f})$.
- Consider the following family of "reference" models:
  - $\mathcal{A}^{\mathrm{ref}}$ is the least squares reference algorithm,
  - (U_k, v) is random design with k features, and noise with level σ^2.
- Recall that $\mathrm{DofF}(\mathcal{A}^{\mathrm{ref}}(U_k, v)) = k$ so long as $\mathrm{rank}(U_k) = k$.
- Thus, for a fitting procedure $\widehat{f} = \mathcal{A}(X, y)$, $\mathrm{DofF}(\widehat{f})$ is also equal to the value of k that satisfy the following relation:

$$\mathrm{OptF}(\mathcal{A}(X, y)) = \mathrm{OptF}(\mathcal{A}^{\mathrm{ref}}(U_k, v)) \qquad \text{(dfF)}$$

Re-interpreting fixed-X degrees of freedom

Fixed-X degrees of freedom is a standard algorithm specific measure of complexity, but no notion of random-X degrees of freedom we know of.

We cast fixed-X degrees of freedom from a different perspective.

- Define fixed-X optimism of $\widehat{f}$ by $\mathrm{OptF}(\widehat{f}) = \mathrm{ErrF}(\widehat{f}) - \mathrm{ErrT}(\widehat{f})$.
- Consider the following family of "reference" models:
  - $\mathcal{A}^{\mathrm{ref}}$ is the least squares reference algorithm,
  - (U_k, v) is random design with k features, and noise with level σ^2.
- Recall that $\mathrm{DofF}(\mathcal{A}^{\mathrm{ref}}(U_k, v)) = k$ so long as $\mathrm{rank}(U_k) = k$.
- Thus, for a fitting procedure $\widehat{f} = \mathcal{A}(X, y)$, $\mathrm{DofF}(\widehat{f})$ is also equal to the value of k that satisfy the following relation:

$$\mathrm{OptF}(\mathcal{A}(X, y)) = \mathrm{OptF}(\mathcal{A}^{\mathrm{ref}}(U_k, v)) \qquad \text{(dfF)}$$

Re-interpreting fixed-X degrees of freedom

Fixed-X degrees of freedom is a standard algorithm specific measure of complexity, but no notion of random-X degrees of freedom we know of.

We cast fixed-X degrees of freedom from a different perspective.

- Define fixed-X optimism of $\widehat{f}$ by $\mathrm{OptF}(\widehat{f}) = \mathrm{ErrF}(\widehat{f}) - \mathrm{ErrT}(\widehat{f})$.
- Consider the following family of "reference" models:
  - $\mathcal{A}^{\mathrm{ref}}$ is the least squares reference algorithm,
  - (U_k, v) is random design with k features, and noise with level σ^2.
- Recall that $\mathrm{DofF}(\mathcal{A}^{\mathrm{ref}}(U_k, v)) = k$ so long as $\mathrm{rank}(U_k) = k$.
- Thus, for a fitting procedure $\widehat{f} = \mathcal{A}(X, y)$, $\mathrm{DofF}(\widehat{f})$ is also equal to the value of k that satisfy the following relation:

$$\mathrm{OptF}(\mathcal{A}(X, y)) = \mathrm{OptF}(\mathcal{A}^{\mathrm{ref}}(U_k, v)) \qquad \text{(dfF)}$$

Emergent random-X degrees of freedom

"Matching optimism" interpretation can be extended to random-X setting and leads to the definition of random-X degrees of freedom.

- Define random-X optimism of $\widehat{f}$ by $\mathrm{OptR}(\widehat{f}) = \mathrm{ErrR}(\widehat{f}) - \mathrm{ErrT}(\widehat{f})$, where $\mathrm{ErrR}(\widehat{f}) = \mathbb{E}[(y_0 - \widehat{f}(x_0))^2]$ is the random-X prediction error.
- We thus define the random-X degrees of freedom, $\mathrm{DofR}(\widehat{f})$, of any predictor $\widehat{f} = \mathcal{A}(X, y)$, as the value of k for which the following relation holds:

$$\mathrm{OptR}(\mathcal{A}(X, y)) = \mathrm{OptR}(\mathcal{A}^{\mathrm{ref}}(U_k, v)) \qquad \text{(dfR, emergent)}$$

Recall here:

- $\mathcal{A}^{\mathrm{ref}}$ is the least squares reference algorithm,
- (U_k, v) is random design with k features, and noise with level σ^2.

We call the measure emergent random-X degrees of freedom.

Emergent random-X degrees of freedom

"Matching optimism" interpretation can be extended to random-X setting and leads to the definition of random-X degrees of freedom.

- Define random-X optimism of $\widehat{f}$ by $\mathrm{OptR}(\widehat{f}) = \mathrm{ErrR}(\widehat{f}) - \mathrm{ErrT}(\widehat{f})$, where $\mathrm{ErrR}(\widehat{f}) = \mathbb{E}[(y_0 - \widehat{f}(x_0))^2]$ is the random-X prediction error.
- We thus define the random-X degrees of freedom, $\mathrm{DofR}(\widehat{f})$, of any predictor $\widehat{f} = \mathcal{A}(X, y)$, as the value of k for which the following relation holds:

$$\mathrm{OptR}(\mathcal{A}(X, y)) = \mathrm{OptR}(\mathcal{A}^{\mathrm{ref}}(U_k, v)) \qquad \text{(dfR, emergent)}$$

Recall here:

- $\mathcal{A}^{\mathrm{ref}}$ is the least squares reference algorithm,
- (U_k, v) is random design with k features, and noise with level σ^2.

We call the measure emergent random-X degrees of freedom.

Emergent random-X degrees of freedom

"Matching optimism" interpretation can be extended to random-X setting and leads to the definition of random-X degrees of freedom.

- Define random-X optimism of $\widehat{f}$ by $\mathrm{OptR}(\widehat{f}) = \mathrm{ErrR}(\widehat{f}) - \mathrm{ErrT}(\widehat{f})$, where $\mathrm{ErrR}(\widehat{f}) = \mathbb{E}[(y_0 - \widehat{f}(x_0))^2]$ is the random-X prediction error.
- We thus define the random-X degrees of freedom, $\mathrm{DofR}(\widehat{f})$, of any predictor $\widehat{f} = \mathcal{A}(X, y)$, as the value of k for which the following relation holds:

$$\mathrm{OptR}(\mathcal{A}(X, y)) = \mathrm{OptR}(\mathcal{A}^{\mathrm{ref}}(U_k, v)) \qquad \text{(dfR, emergent)}$$

Recall here:

- $\mathcal{A}^{\mathrm{ref}}$ is the least squares reference algorithm,
- (U_k, v) is random design with k features, and noise with level σ^2.

We call the measure emergent random-X degrees of freedom.

Emergent random-X degrees of freedom

"Matching optimism" interpretation can be extended to random-X setting and leads to the definition of random-X degrees of freedom.

- Define random-X optimism of $\widehat{f}$ by $\mathrm{OptR}(\widehat{f}) = \mathrm{ErrR}(\widehat{f}) - \mathrm{ErrT}(\widehat{f})$, where $\mathrm{ErrR}(\widehat{f}) = \mathbb{E}[(y_0 - \widehat{f}(x_0))^2]$ is the random-X prediction error.
- We thus define the random-X degrees of freedom, $\mathrm{DofR}(\widehat{f})$, of any predictor $\widehat{f} = \mathcal{A}(X, y)$, as the value of k for which the following relation holds:

$$\mathrm{OptR}(\mathcal{A}(X, y)) = \mathrm{OptR}(\mathcal{A}^{\mathrm{ref}}(U_k, v)) \qquad \text{(dfR, emergent)}$$

Recall here:

- $\mathcal{A}^{\mathrm{ref}}$ is the least squares reference algorithm,
- (U_k, v) is random design with k features, and noise with level σ^2.

We call the measure emergent random-X degrees of freedom.

Emergent random-X degrees of freedom

"Matching optimism" interpretation can be extended to random-X setting and leads to the definition of random-X degrees of freedom.

- Define random-X optimism of $\widehat{f}$ by $\mathrm{OptR}(\widehat{f}) = \mathrm{ErrR}(\widehat{f}) - \mathrm{ErrT}(\widehat{f})$, where $\mathrm{ErrR}(\widehat{f}) = \mathbb{E}[(y_0 - \widehat{f}(x_0))^2]$ is the random-X prediction error.
- We thus define the random-X degrees of freedom, $\mathrm{DofR}(\widehat{f})$, of any predictor $\widehat{f} = \mathcal{A}(X, y)$, as the value of k for which the following relation holds:

$$\mathrm{OptR}(\mathcal{A}(X, y)) = \mathrm{OptR}(\mathcal{A}^{\mathrm{ref}}(U_k, v)) \qquad \text{(dfR, emergent)}$$

Recall here:

- $\mathcal{A}^{\mathrm{ref}}$ is the least squares reference algorithm,
- (U_k, v) is random design with k features, and noise with level σ^2.

We call the measure emergent random-X degrees of freedom.

Intrisic random-X degrees of freedom

- The emergent random-X degrees of freedom, $\mathrm{DofR}(\widehat{f})$, depends of both the the predictor $\widehat{f}$ and the underlying regression function f.
- When matching optimisms, the observed random-X optimism of $\widehat{f}$ consists of bias, which may inflate the degrees of freedom.
- We thus also define intrinsic random-X degrees of freedom, denoted by DofR^i, as the k for which the following relation holds:

$$\mathrm{OptR}(\mathcal{A}(X, v)) = \mathrm{OptR}(\mathcal{A}^{\mathrm{ref}}(U_k, v)) \qquad \text{(dfR, intrinsic)}$$

The intrinsic random-X degrees of freedom measures the inherent complexity of the predictor $\widehat{f}$ in terms of overfitting to “pure noise”.

Intrisic random-X degrees of freedom

- The emergent random-X degrees of freedom, $\mathrm{DofR}(\widehat{f})$, depends of both the the predictor $\widehat{f}$ and the underlying regression function f.
- When matching optimisms, the observed random-X optimism of $\widehat{f}$ consists of bias, which may inflate the degrees of freedom.
- We thus also define intrinsic random-X degrees of freedom, denoted by DofR^i, as the k for which the following relation holds:

$$\mathrm{OptR}(\mathcal{A}(X, v)) = \mathrm{OptR}(\mathcal{A}^{\mathrm{ref}}(U_k, v)) \qquad \text{(dfR, intrinsic)}$$

The intrinsic random-X degrees of freedom measures the inherent complexity of the predictor $\widehat{f}$ in terms of overfitting to "pure noise".

Intrisic random-X degrees of freedom

- The emergent random-X degrees of freedom, $\mathrm{DofR}(\widehat{f})$, depends of both the the predictor $\widehat{f}$ and the underlying regression function f.
- When matching optimisms, the observed random-X optimism of $\widehat{f}$ consists of bias, which may inflate the degrees of freedom.
- We thus also define intrinsic random-X degrees of freedom, denoted by DofR^{i}, as the k for which the following relation holds:

$$\mathrm{OptR}(\mathcal{A}(X, v)) = \mathrm{OptR}(\mathcal{A}^{\mathrm{ref}}(U_k, v)) \qquad \text{(dfR, intrinsic)}$$

The intrinsic random-X degrees of freedom measures the inherent complexity of the predictor $\widehat{f}$ in terms of overfitting to “pure noise”.

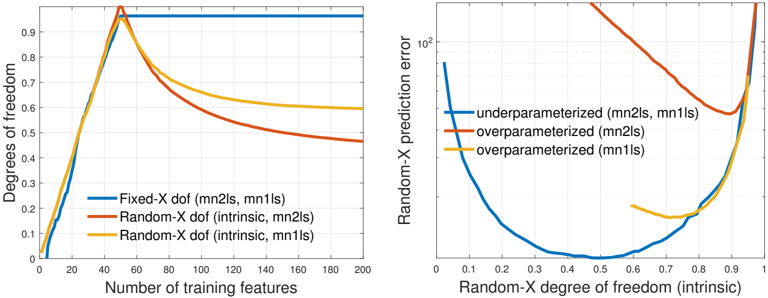
Intrisic random-X degrees of freedom

- The emergent random-X degrees of freedom, $\mathrm{DofR}(\widehat{f})$, depends of both the the predictor $\widehat{f}$ and the underlying regression function f.
- When matching optimisms, the observed random-X optimism of $\widehat{f}$ consists of bias, which may inflate the degrees of freedom.
- We thus also define intrinsic random-X degrees of freedom, denoted by DofR^i, as the k for which the following relation holds:

$$\mathrm{OptR}(\mathcal{A}(X, v)) = \mathrm{OptR}(\mathcal{A}^{\mathrm{ref}}(U_k, v)) \qquad \text{(dfR, intrinsic)}$$

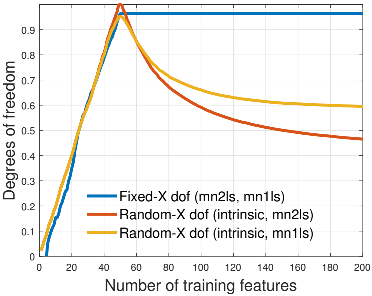
The intrinsic random-X degrees of freedom measures the inherent complexity of the predictor $\widehat{f}$ in terms of overfitting to “pure noise”.

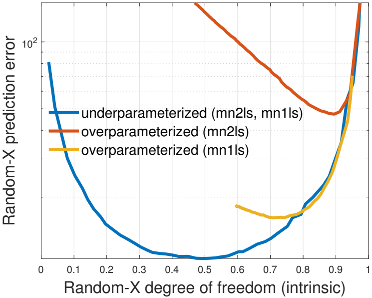
Random-X degrees of freedom illustration



- Fixed data with $n = 50$ and response non-linear in $p = 200$ features
- Model class: estimators fitted on nested subsets of 1 to 200 feaures
- Fixed-X: increase then constant; random-X: increase then decrease
- Underparameterized: U-curve; overparameterized: also U-curve!
- Punchline: reparameterize overparameterized to underparameterized

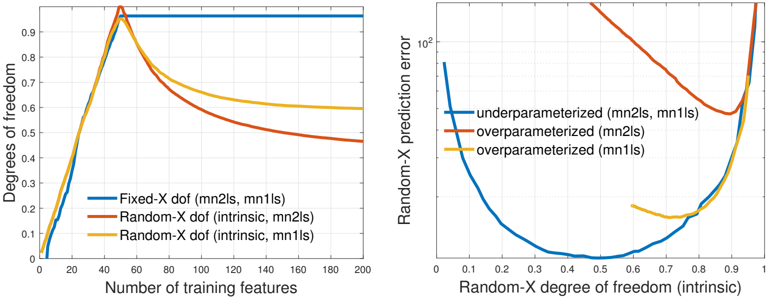
Random-X degrees of freedom illustration





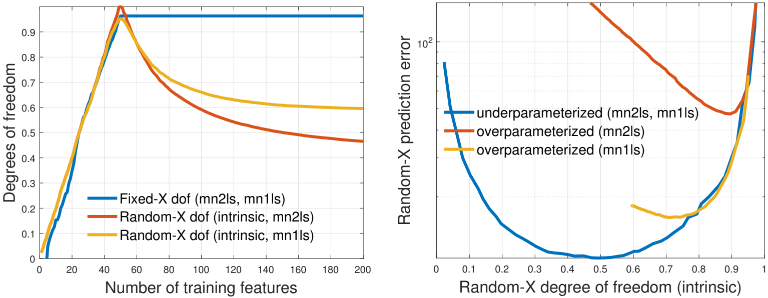
- Fixed data with $n = 50$ and response non-linear in $p = 200$ features
- Model class: estimators fitted on nested subsets of 1 to 200 feaures
- Fixed-X: increase then constant; random-X: increase then decrease
- Underparameterized: U-curve; overparameterized: also U-curve!
- Punchline: reparameterize overparameterized to underparameterized

Random-X degrees of freedom illustration



- Fixed data with $n = 50$ and response non-linear in $p = 200$ features
- Model class: estimators fitted on nested subsets of 1 to 200 feaures
- Fixed-X: increase then constant; random-X: increase then decrease
- Underparameterized: U-curve; overparameterized: also U-curve!
- Punchline: reparameterize overparameterized to underparameterized

Random-X degrees of freedom illustration



- Fixed data with $n = 50$ and response non-linear in $p = 200$ features
- Model class: estimators fitted on nested subsets of 1 to 200 feaures
- Fixed-X: increase then constant; random-X: increase then decrease
- Underparameterized: U-curve; overparameterized: also U-curve!
- Punchline: reparameterize overparameterized to underparameterized

Discussion and future directions

A high-level view of the work:

- Suppose we are given a family of models for which we want a complexity measure under a specific error metric.
- Construct a family of "reference" models spanning same optimisms.
- Find the model in the reference family that is closest to the observed optimism. Declare complexity as complexity of that reference model.

Key relation:

$$\mathrm{OptR}(\widehat{f}) = \mathrm{OptR}(\widehat{f}^{\mathrm{ref}})$$

Future directions:

- Attribute total complexity to various components: bias, variance, covariate shift, etc.
- Other error metrics beyond squared error

Discussion and future directions

A high-level view of the work:

- Suppose we are given a family of models for which we want a complexity measure under a specific error metric.
- Construct a family of "reference" models spanning same optimisms.
- Find the model in the reference family that is closest to the observed optimism. Declare complexity as complexity of that reference model.

Key relation:

$$\mathrm{OptR}(\widehat{f}) = \mathrm{OptR}(\widehat{f}^{\mathrm{ref}})$$

Future directions:

- Attribute total complexity to various components: bias, variance, covariate shift, etc.
- Other error metrics beyond squared error

Discussion and future directions

A high-level view of the work:

- Suppose we are given a family of models for which we want a complexity measure under a specific error metric.
- Construct a family of "reference" models spanning same optimisms.
- Find the model in the reference family that is closest to the observed optimism. Declare complexity as complexity of that reference model.

Key relation:

$$\mathrm{OptR}(\widehat{f}) = \mathrm{OptR}(\widehat{f}^{\mathrm{ref}})$$

Future directions:

- Attribute total complexity to various components: bias, variance, covariate shift, etc.
- Other error metrics beyond squared error

Discussion and future directions

A high-level view of the work:

- Suppose we are given a family of models for which we want a complexity measure under a specific error metric.
- Construct a family of "reference" models spanning same optimisms.
- Find the model in the reference family that is closest to the observed optimism. Declare complexity as complexity of that reference model.

Key relation:

$$\mathrm{OptR}(\widehat{f}) = \mathrm{OptR}(\widehat{f}^{\mathrm{ref}})$$

Future directions:

- Attribute total complexity to various components: bias, variance, covariate shift, etc.
- Other error metrics beyond squared error

Discussion and future directions

A high-level view of the work:

- Suppose we are given a family of models for which we want a complexity measure under a specific error metric.
- Construct a family of "reference" models spanning same optimisms.
- Find the model in the reference family that is closest to the observed optimism. Declare complexity as complexity of that reference model.

Key relation:

$$\mathrm{OptR}(\widehat{f}) = \mathrm{OptR}(\widehat{f}^{\mathrm{ref}})$$

Future directions:

- Attribute total complexity to various components: bias, variance, covariate shift, etc.
- Other error metrics beyond squared error

Discussion and future directions

A high-level view of the work:

- Suppose we are given a family of models for which we want a complexity measure under a specific error metric.
- Construct a family of "reference" models spanning same optimisms.
- Find the model in the reference family that is closest to the observed optimism. Declare complexity as complexity of that reference model.

Key relation:

$$\mathrm{OptR}(\widehat{f}) = \mathrm{OptR}(\widehat{f}^{\mathrm{ref}})$$

Future directions:

- Attribute total complexity to various components: bias, variance, covariate shift, etc.
- Other error metrics beyond squared error

Discussion and future directions

A high-level view of the work:

- Suppose we are given a family of models for which we want a complexity measure under a specific error metric.
- Construct a family of "reference" models spanning same optimisms.
- Find the model in the reference family that is closest to the observed optimism. Declare complexity as complexity of that reference model.

Key relation:

$$\mathrm{OptR}(\widehat{f}) = \mathrm{OptR}(\widehat{f}^{\mathrm{ref}})$$

Future directions:

- Attribute total complexity to various components: bias, variance, covariate shift, etc.
- Other error metrics beyond squared error

Discussion and future directions

A high-level view of the work:

- Suppose we are given a family of models for which we want a complexity measure under a specific error metric.
- Construct a family of "reference" models spanning same optimisms.
- Find the model in the reference family that is closest to the observed optimism. Declare complexity as complexity of that reference model.

Key relation:

$$\mathrm{OptR}(\widehat{f}) = \mathrm{OptR}(\widehat{f}^{\mathrm{ref}})$$

Future directions:

- Attribute total complexity to various components: bias, variance, covariate shift, etc.
- Other error metrics beyond squared error

Outline

Motivating thesis questions with take-aways

We studied three operational aspects of overparameterized learning:
1) cross-validation, 2) risk monotonization, 3) model complexity.

1. Cross-validation still works in the overparameterized regime, especially when optimal regularization and train error can be zero for ridge regression through analytic continuation.

2. It is possible to modify any given prediction procedure to mitigate double descent behavior and achieve a monotonic risk behavior through subsampling and cross-validation.

3. There is a principled measure of model complexity in general for overparameterized models in the form of random-X degrees of freedom.

Motivating thesis questions with take-aways

We studied three operational aspects of overparameterized learning:
1) cross-validation, 2) risk monotonization, 3) model complexity.

1. Cross-validation still works in the overparameterized regime, especially when optimal regularization and train error can be zero for ridge regression through analytic continuation.

2. It is possible to modify any given prediction procedure to mitigate double descent behavior and achieve a monotonic risk behavior through subsampling and cross-validation.

3. There is a principled measure of model complexity in general for overparameterized models in the form of random-X degrees of freedom.

Motivating thesis questions with take-aways

We studied three operational aspects of overparameterized learning:
1) cross-validation, 2) risk monotonization, 3) model complexity.

1. Cross-validation still works in the overparameterized regime, especially when optimal regularization and train error can be zero for ridge regression through analytic continuation.

2. It is possible to modify any given prediction procedure to mitigate double descent behavior and achieve a monotonic risk behavior through subsampling and cross-validation.

3. There is a principled measure of model complexity in general for overparameterized models in the form of random-X degrees of freedom.

Motivating thesis questions with take-aways

We studied three operational aspects of overparameterized learning:
1) cross-validation, 2) risk monotonization, 3) model complexity.

1. Cross-validation still works in the overparameterized regime, especially when optimal regularization and train error can be zero for ridge regression through analytic continuation.

2. It is possible to modify any given prediction procedure to mitigate double descent behavior and achieve a monotonic risk behavior through subsampling and cross-validation.

3. There is a principled measure of model complexity in general for overparameterized models in the form of random-X degrees of freedom.

Thanks for listening!

Questions/comments/thoughts?

BIG THANKS!

- Ryan
- Committee: Ale, Arun, Yuting, Arian
- Collaborators
- Faculty
- Staff
- Students
- Funding agency

BIG THANKS!

- Ryan
- Committee: Ale, Arun, Yuting, Arian
- Collaborators
- Faculty
- Staff
- Students
- Funding agency

BIG THANKS!

- Ryan
- Committee: Ale, Arun, Yuting, Arian
- Collaborators
- Faculty
- Staff
- Students
- Funding agency

BIG THANKS!

- Ryan
- Committee: Ale, Arun, Yuting, Arian
- Collaborators
- Faculty
- Staff
- Students
- Funding agency

BIG THANKS!

- Ryan
- Committee: Ale, Arun, Yuting, Arian
- Collaborators
- Faculty
- Staff
- Students
- Funding agency

BIG THANKS!

- Ryan
- Committee: Ale, Arun, Yuting, Arian
- Collaborators
- Faculty
- Staff
- Students
- Funding agency

BIG THANKS!

- Ryan
- Committee: Ale, Arun, Yuting, Arian
- Collaborators
- Faculty
- Staff
- Students
- Funding agency