Personalized Treatment: Sounds heavenly, but where on Earth did they find *my* guinea pigs?

Xiao-Li Meng, Harvard University

- Meng (2014) A trio of inference problems that could win you a Nobel Prize in statistics (if you help fund it). *In Past, Present and Future of Statistics*. COPSS 50th Anniversary Vol. (Ed: Lin et. al)
- Liu and Meng (2014). A fruitful resolution to Simpson's paradox via multi-resolution inference. *The American Statistician*, 68: 17-29.
- Liu and Meng (2016). There is individualized treatment. Why not individualized inference?. *Annual Review of Statistics and Its Application*, 3, 79-111.

A Multi-resolution Theory for Approximating Infinite-p-Zero-n: Transitional Inference, Individualized Predictions, and a World without Bias-variance Trade-off (*JASA* (2021) special issue on "Precision medicine and individualized policy discovery")

Xinran Li and Xiao-Li Meng

- Meng (2014) **A trio of inference problems that could win you a Nobel Prize in statistics (if you help fund it).** In *Past, Present and Future of Statistics. COPSS 50th Anniversary Volume* (Ed: Lin et. al.)
- Liu and Meng (2014). **A fruitful resolution to Simpson's paradox via multi-resolution inference.** *The American Statistician*, 68: 17-29.
- Liu and Meng (2016). **There is individualized treatment. Why not individualized inference?.** *Annual Review of Statistics and Its Application*, 3, 79-111.

A Big Picture about Data Science

A Big Picture about Data Science

"Whereas previous generations were in possession of data about general phenomena, we are now in possession of data about specific phenomena. For example, in genomics we have data about each individual gene, in astronomy we have data about each region of the sky, in medicine we have data about each tumor, and in social science we have data about individual humans. Our era is about 'data' and about 'specific context.' In that sense 'data science' is an appropriate and useful terminology for capturing current trends."

Ani Adhikari, John DeNero, and Michael Jordan
(*Harvard Data Science Review*, Winter, 2021)

Promises of Big Data

Treatment for you based only on data from people like you.

Promises of Big Data

Treatment for you based only on data from people like you.

BUT: No one is perfectly like you.

Promises of Big Data

Treatment for you based only on data from people like you.

BUT: No one is perfectly like you.

Relevance for COVID-19: Population Risk vs Individual Risk

What are the p and n for individualized treatments?

What are the p and n for individualized treatments?

- **Potentially Infinitely many attributes:** $p = \infty$
- **Each of us is unique: direct learning sample size** $n = 0$**.**

What are the p and n for individualized treatments?

- **Potentially Infinitely many attributes:** $p = \infty$
- **Each of us is unique: direct learning sample size** $n = 0$**.**

- **Moving from population-level "soft matching" to individual-level "hard matching" – is that possible?**

What are the p and n for individualized treatments?

- **Potentially Infinitely many attributes: $p = \infty$**
- **Each of us is unique: direct learning sample size $n = 0$.**

- **Moving from population-level "soft matching" to individual-level "hard matching" – is that possible?**

Ultimate bias-variance trade-off

Relevance vs Robustness

It's all Greek (or philosophy) to me ...

- "**Transition to the similar**"

It's all Greek (or philosophy) to me ...

- "**Transition to the similar**"

  Aelius Galenus (129-210 AD)

It's all Greek (or philosophy) to me ...

- "**Transition to the similar**"

  Aelius Galenus (129-210 AD)

- **Transitional inference**: *an empiricism concept* (Hankinson 1987, 1995)

It's all Greek (or philosophy) to me ...

- "**Transition to the similar**"

  Aelius Galenus (129-210 AD)

- **Transitional inference**: *an empiricism concept* (Hankinson 1987, 1995)

- "**In cases in which there is no history, or in which there is none of sufficient similarity, there is not much hope.**

It's all Greek (or philosophy) to me ...

- "**Transition to the similar**"
  Aelius Galenus (129-210 AD)
- **Transitional inference**: *an empiricism concept* (Hankinson 1987, 1995)

- "**In cases in which there is no history, or in which there is none of sufficient similarity, there is not much hope. And the same thing is true in the case of transference of one remedy from one ailment to another similar to it: one has a greater or smaller basis for expectation of success in proportion to the increase or decrease in similarity of the ailment, whether or not history is involved.**

It's all Greek (or philosophy) to me ...

- "**Transition to the similar**"
  Aelius Galenus (129-210 AD)
- **Transitional inference**: *an empiricism concept* (Hankinson 1987, 1995)

- "**In cases in which there is no history, or in which there is none of sufficient similarity, there is not much hope. And the same thing is true in the case of transference of one remedy from one ailment to another similar to it: one has a greater or smaller basis for expectation of success in proportion to the increase or decrease in similarity of the ailment, whether or not history is involved. And the same goes for the transference from one part of the body to another part: expectation of success varies in direct proportion to the similarity.**"

Inference/Prediction as Approximation

- **What Do We Want to Know?**: Your response to treatment t, $Y_t(☺^*)$, for $t = A$ and $t = B$.

Inference/Prediction as Approximation

- **What Do We Want to Know?**: Your response to treatment t, $Y_t(☺^*)$, for $t = A$ and $t = B$.
- **What Do We Know?**: Either $Y_A(☺)$ or $Y_B(☺)$ (but not both) for some other people, $☺ \neq ☺^*$.

Inference/Prediction as Approximation

- **What Do We Want to Know?**: Your response to treatment t, $Y_t(☺^*)$, for $t = A$ and $t = B$.
- **What Do We Know?**: Either $Y_A(☺)$ or $Y_B(☺)$ (but not both) for some other people, $☺ \neq ☺^*$.
- **Strategy**: Must construct population of "relevant" individuals with which to approximate **YOU**.

Resolution of the Approximating Population

- The ☺* **relevant subpopulation** is

$$\Omega_C(☺^*) = \{☺ : C(☺) = C(☺^*)\}.$$

where C is a set of intrinsic (pre-treatment) characteristics.

Resolution of the Approximating Population

- The ☺^* **relevant subpopulation** is

$$\Omega_C(\text{☺}^*) = \{\text{☺} : C(\text{☺}) = C(\text{☺}^*)\}.$$

  where C is a set of intrinsic (pre-treatment) characteristics.

- $R = \dim(C)$ is the **primary resolution**.

Resolution of the Approximating Population

- The ☺* **relevant subpopulation** is

$$\Omega_C(☺^*) = \{☺ : C(☺) = C(☺^*)\}.$$

  where C is a set of intrinsic (pre-treatment) characteristics.

- R =dim(C) is the **primary resolution**.
- Inspired by **multi-resolution** wavelets formulation: match on "signals" (low resolutions) and ignore "noises" (high resolutions)

Resolution of the Approximating Population

- The ☺* **relevant subpopulation** is

$$\Omega_C(☺^*) = \{☺ : C(☺) = C(☺^*)\}.$$

  where C is a set of intrinsic (pre-treatment) characteristics.

- R =dim(C) is the **primary resolution**.
- Inspired by **multi-resolution** wavelets formulation: match on "signals" (low resolutions) and ignore "noises" (high resolutions)

Resolution of the Approximating Population

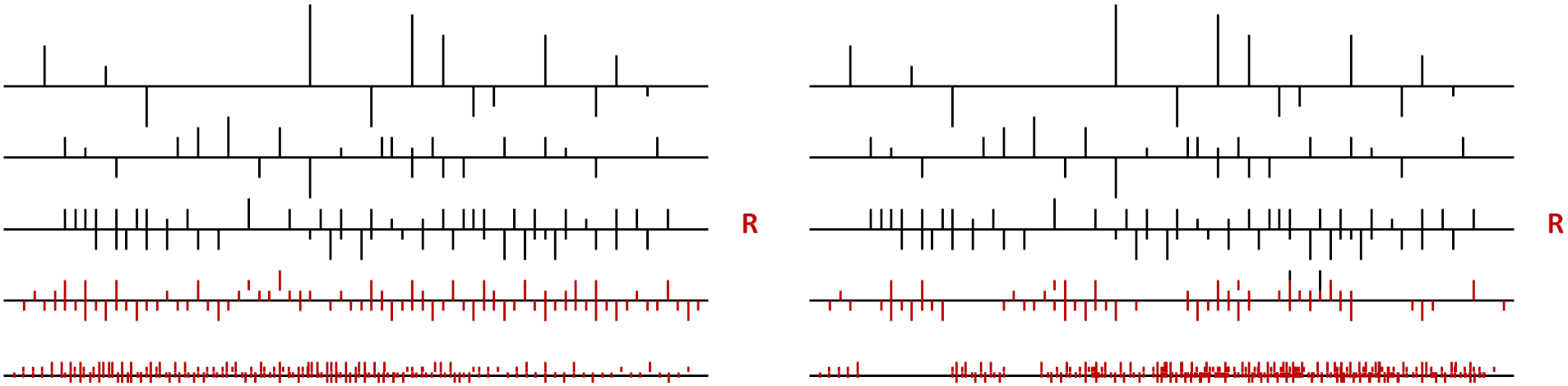
- The ☺* **relevant subpopulation** is

$$\Omega_C(☺^*) = \{☺ : C(☺) = C(☺^*)\}.$$

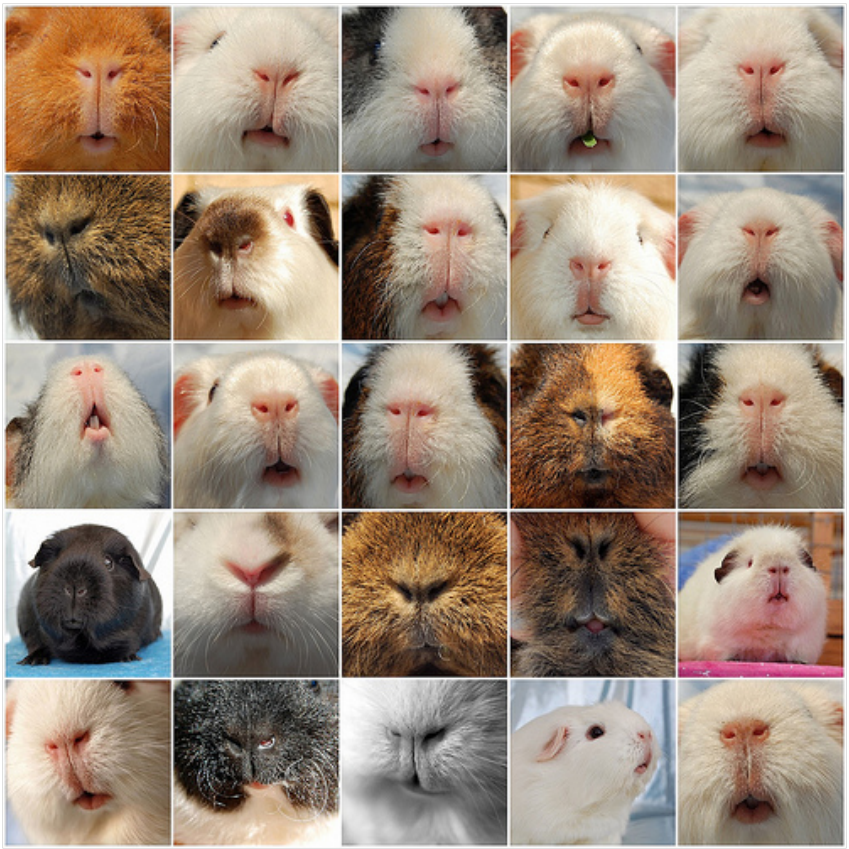
  where C is a set of intrinsic (pre-treatment) characteristics.

- R =dim(C) is the **primary resolution**.
- Inspired by **multi-resolution** wavelets formulation: match on "signals" (low resolutions) and ignore "noises" (high resolutions)

YOU A Relevant Individual

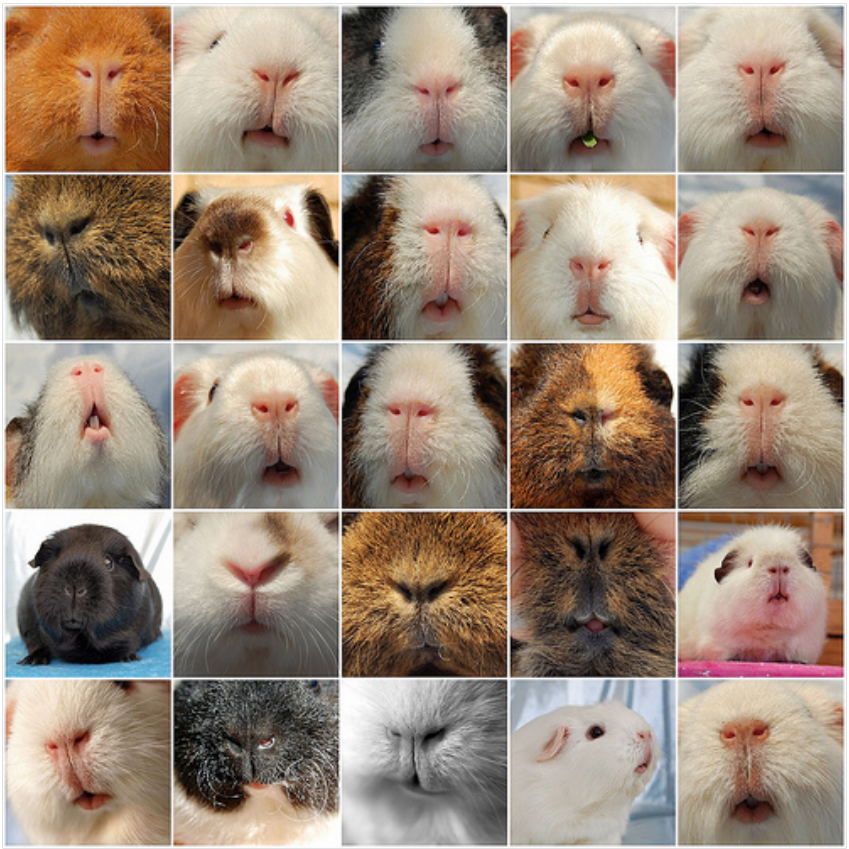


Population Resolution



A Multi-Resolution View of Big Data

Population Resolution



Individual Resolution

How to Capture Resolution?

- Y lives on the same probability space as an information filtration $\{\mathcal{F}_r, r = 0, 1, ..., \}$, e.g., $\mathcal{F}_r = \sigma(X_0, X_1, \ldots, X_r)$.

How to Capture Resolution?

- Y lives on the same probability space as an information filtration $\{\mathcal{F}_r, r = 0, 1, ..., \}$, e.g., $\mathcal{F}_r = \sigma(X_0, X_1, \ldots, X_r)$.

- r is the index of resolution.

How to Capture Resolution?

- Y lives on the same probability space as an information filtration $\{\mathcal{F}_r, r = 0, 1, ..., \}$, e.g., $\mathcal{F}_r = \sigma(X_0, X_1, \ldots, X_r)$.
- r is the index of resolution.

Resolution Specific Signal and Noise

$$\underbrace{\mu_r = \mathrm{E}[Y|\mathcal{F}_r]}_{\substack{\text{signal at} \\ \text{resolution } r}} \quad \text{and} \quad \underbrace{\sigma_r^2 = \mathrm{V}[Y|\mathcal{F}_r]}_{\substack{\text{noise at} \\ \text{resolution } r}}$$

Conditional ANOVA Decomposition (Pythagoras Theorem): for $s > r$

$$\sigma_r^2 = \mathsf{E}[\sigma_s^2|\mathcal{F}_r] + \mathsf{E}[(\mu_s - \mu_r)^2|\mathcal{F}_r]$$

Conditional ANOVA Decomposition (Pythagoras Theorem): for $s > r$

$$\sigma_r^2 = \mathsf{E}[\sigma_s^2|\mathcal{F}_r] + \mathsf{E}[(\mu_s - \mu_r)^2|\mathcal{F}_r]$$

- $(\mu_s - \mu_r)^2$: magnitude of signal at resolution s, not modelled at resolution r.

Conditional ANOVA Decomposition (Pythagoras Theorem): for $s > r$

$$\sigma_r^2 = \mathsf{E}[\sigma_s^2|\mathcal{F}_r] + \mathsf{E}[(\mu_s - \mu_r)^2|\mathcal{F}_r]$$

- $(\mu_s - \mu_r)^2$: magnitude of signal at resolution s, not modelled at resolution r.

Modelling "Individualized Treatment" as $r \to \infty$

$$\sigma_r^2 = \underbrace{\mathsf{E}[\sigma_\infty^2|\mathcal{F}_r]}_{\text{intrinsic variance}} + \underbrace{\sum_{i=r}^{\infty} \mathsf{E}[(\mu_{i+1} - \mu_i)^2|\mathcal{F}_r]}_{\text{total resolution bias}}$$

KEY: No such thing called "Variance" except at resolution ∞

Conditional ANOVA Decomposition (Pythagoras Theorem): for $s > r$

$$\sigma_r^2 = \mathsf{E}[\sigma_s^2|\mathcal{F}_r] + \mathsf{E}[(\mu_s - \mu_r)^2|\mathcal{F}_r]$$

- $(\mu_s - \mu_r)^2$: magnitude of signal at resolution s, not modelled at resolution r.

Modelling "Individualized Treatment" as $r \to \infty$

$$\sigma_r^2 = \underbrace{\mathsf{E}[\sigma_\infty^2|\mathcal{F}_r]}_{\text{intrinsic variance}} + \underbrace{\sum_{i=r}^{\infty} \mathsf{E}[(\mu_{i+1} - \mu_i)^2|\mathcal{F}_r]}_{\text{total resolution bias}}$$

KEY: No such thing called "Variance" except at resolution ∞

(1) If we believe the world is stochastic, then $\sigma_\infty^2 > 0$.

Conditional ANOVA Decomposition (Pythagoras Theorem): for $s > r$

$$\sigma_r^2 = \mathrm{E}[\sigma_s^2|\mathcal{F}_r] + \mathrm{E}[(\mu_s - \mu_r)^2|\mathcal{F}_r]$$

- $(\mu_s - \mu_r)^2$: magnitude of signal at resolution s, not modelled at resolution r.

Modelling "Individualized Treatment" as $r \to \infty$

$$\sigma_r^2 = \underbrace{\mathrm{E}[\sigma_\infty^2|\mathcal{F}_r]}_{\text{intrinsic variance}} + \underbrace{\sum_{i=r}^{\infty} \mathrm{E}[(\mu_{i+1} - \mu_i)^2|\mathcal{F}_r]}_{\text{total resolution bias}}$$

KEY: No such thing called "Variance" except at resolution ∞

(1) If we believe the world is stochastic, then $\sigma_\infty^2 > 0$.

(2) If we believe the world is deterministic/chaotic, then $\sigma_\infty^2 = 0$.

Conditional ANOVA Decomposition (Pythagoras Theorem): for $s > r$

$$\sigma_r^2 = \mathrm{E}[\sigma_s^2|\mathcal{F}_r] + \mathrm{E}[(\mu_s - \mu_r)^2|\mathcal{F}_r]$$

- $(\mu_s - \mu_r)^2$: magnitude of signal at resolution s, not modelled at resolution r.

Modelling "Individualized Treatment" as $r \to \infty$

$$\sigma_r^2 = \underbrace{\mathrm{E}[\sigma_\infty^2|\mathcal{F}_r]}_{\text{intrinsic variance}} + \underbrace{\sum_{i=r}^{\infty} \mathrm{E}[(\mu_{i+1} - \mu_i)^2|\mathcal{F}_r]}_{\text{total resolution bias}}$$

KEY: No such thing called "Variance" except at resolution ∞

(1) If we believe the world is stochastic, then $\sigma_\infty^2 > 0$.

(2) If we believe the world is deterministic/chaotic, then $\sigma_\infty^2 = 0$.

(3) There is no general empirical test to distinguish between (1) and (2), with any finite amount of data.

Shifting from $n \to 0$ to $r \to \infty$ ...

- Without direct data, we estimate Y_{me} using $\hat{\mu}_r$ with $r = \tilde{r}$

Errors in Indirect-Data Prediction

$$Y_{me} - \hat{\mu}_{\tilde{r}} = \underbrace{(\mu_{\tilde{r}} - \hat{\mu}_{\tilde{r}})}_{\substack{\text{Estimation Error} \\ \textbf{“Variance”}}} + \underbrace{(\mu_\infty - \mu_{\tilde{r}})}_{\substack{\text{Resolution Error} \\ \textbf{Bias}}} + \underbrace{(Y_{me} - \mu_\infty)}_{\substack{\text{Intrinsic Error} \\ \textbf{Intrinsic Variance}}}$$

Shifting from $n \to 0$ to $r \to \infty$ ...

- Without direct data, we estimate Y_{me} using $\hat{\mu}_r$ with $r = \tilde{r}$

Errors in Indirect-Data Prediction

$$Y_{me} - \hat{\mu}_{\tilde{r}} = \underbrace{(\mu_{\tilde{r}} - \hat{\mu}_{\tilde{r}})}_{\substack{\text{Estimation Error} \\ \textbf{“Variance”}}} + \underbrace{(\mu_\infty - \mu_{\tilde{r}})}_{\substack{\text{Resolution Error} \\ \textbf{Bias}}} + \underbrace{(Y_{me} - \mu_\infty)}_{\substack{\text{Intrinsic Error} \\ \textbf{Intrinsic Variance}}}$$

- Double accent: “hat” for estimation and “tilde” for selection.

Shifting from $n \to 0$ to $r \to \infty$ ...

- Without direct data, we estimate Y_{me} using $\hat{\mu}_r$ with $r = \tilde{r}$

Errors in Indirect-Data Prediction

$$Y_{me} - \hat{\mu}_{\tilde{r}} = \underbrace{(\mu_{\tilde{r}} - \hat{\mu}_{\tilde{r}})}_{\substack{\text{Estimation Error} \\ \textbf{“Variance”}}} + \underbrace{(\mu_\infty - \mu_{\tilde{r}})}_{\substack{\text{Resolution Error} \\ \textbf{Bias}}} + \underbrace{(Y_{me} - \mu_\infty)}_{\substack{\text{Intrinsic Error} \\ \textbf{Intrinsic Variance}}}$$

- Double accent: “hat” for estimation and “tilde” for selection.
- How big should $\tilde{r}$ be? **A Holy Grail of Statistical Inference and Prediction**.

Shifting from $n \to 0$ to $r \to \infty$ ...

- Without direct data, we estimate Y_{me} using $\hat{\mu}_r$ with $r = \tilde{r}$

Errors in Indirect-Data Prediction

$$Y_{me} - \hat{\mu}_{\tilde{r}} = \underbrace{(\mu_{\tilde{r}} - \hat{\mu}_{\tilde{r}})}_{\substack{\text{Estimation Error} \\ \textbf{“Variance”}}} + \underbrace{(\mu_\infty - \mu_{\tilde{r}})}_{\substack{\text{Resolution Error} \\ \textbf{Bias}}} + \underbrace{(Y_{me} - \mu_\infty)}_{\substack{\text{Intrinsic Error} \\ \textbf{Intrinsic Variance}}}$$

- Double accent: “hat” for estimation and “tilde” for selection.
- How big should $\tilde{r}$ be? **A Holy Grail of Statistical Inference and Prediction**.
- Equivalent to bias-variance trade-off, but the resolution framework leads to some surprises.

Shifting from $n \to 0$ to $r \to \infty$ ...

- Without direct data, we estimate Y_{me} using $\hat{\mu}_r$ with $r = \tilde{r}$

Errors in Indirect-Data Prediction

$$Y_{me} - \hat{\mu}_{\tilde{r}} = \underbrace{(\mu_{\tilde{r}} - \hat{\mu}_{\tilde{r}})}_{\substack{\text{Estimation Error} \\ \textbf{“Variance”}}} + \underbrace{(\mu_\infty - \mu_{\tilde{r}})}_{\substack{\text{Resolution Error} \\ \textbf{Bias}}} + \underbrace{(Y_{me} - \mu_\infty)}_{\substack{\text{Intrinsic Error} \\ \textbf{Intrinsic Variance}}}$$

- Double accent: “hat” for estimation and “tilde” for selection.
- How big should $\tilde{r}$ be? **A Holy Grail of Statistical Inference and Prediction**.
- Equivalent to bias-variance trade-off, but the resolution framework leads to some surprises.
- It has the same mathematical setup as **sieve method for non-parametric estimation**.

Linear Model with Potentially Infinitely Many Predictors

- Assume both target and training populations are generated from

$$Y = \sum_{r=0}^{\infty} \beta_r X_r + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \tau^2), \quad \epsilon \perp\!\!\!\perp \vec{\boldsymbol{X}}_\infty, \quad \mathrm{V}(Y) < \infty.$$

$$X_0 = 1, \quad \{X_1, X_2, \ldots\} \text{ are jointly normal distributed,}$$

Linear Model with Potentially Infinitely Many Predictors

- Assume both target and training populations are generated from

$$Y = \sum_{r=0}^{\infty} \beta_r X_r + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \tau^2), \quad \epsilon \perp\!\!\!\perp \vec{\boldsymbol{X}}_\infty, \quad \mathrm{V}(Y) < \infty.$$

$$X_0 = 1, \quad \{X_1, X_2, \ldots\} \text{ are jointly normal distributed,}$$

- The total prediction error under OLS has three parts:

Linear Model with Potentially Infinitely Many Predictors

- Assume both target and training populations are generated from

$$Y = \sum_{r=0}^{\infty} \beta_r X_r + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \tau^2), \quad \epsilon \perp\!\!\!\perp \vec{\boldsymbol{X}}_\infty, \quad \mathrm{V}(Y) < \infty.$$

$$X_0 = 1, \quad \{X_1, X_2, \ldots\} \text{ are jointly normal distributed,}$$

- The total prediction error under OLS has three parts:
  - τ^2: intrinsic error

Linear Model with Potentially Infinitely Many Predictors

- Assume both target and training populations are generated from

$$Y = \sum_{r=0}^{\infty} \beta_r X_r + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \tau^2), \quad \epsilon \perp\!\!\!\perp \vec{\boldsymbol{X}}_\infty, \quad \mathrm{V}(Y) < \infty.$$

$$X_0 = 1, \quad \{X_1, X_2, \ldots\} \text{ are jointly normal distributed,}$$

- The total prediction error under OLS has three parts:
  - τ^2: intrinsic error
  - $A(r) = \sum_{k=r+1}^{\infty} \Delta_k^2$: approx error, $\Delta_k^2 = V(Y|\vec{\boldsymbol{X}}_{r-1}) - V(Y|\vec{\boldsymbol{X}}_r)$

Linear Model with Potentially Infinitely Many Predictors

- Assume both target and training populations are generated from

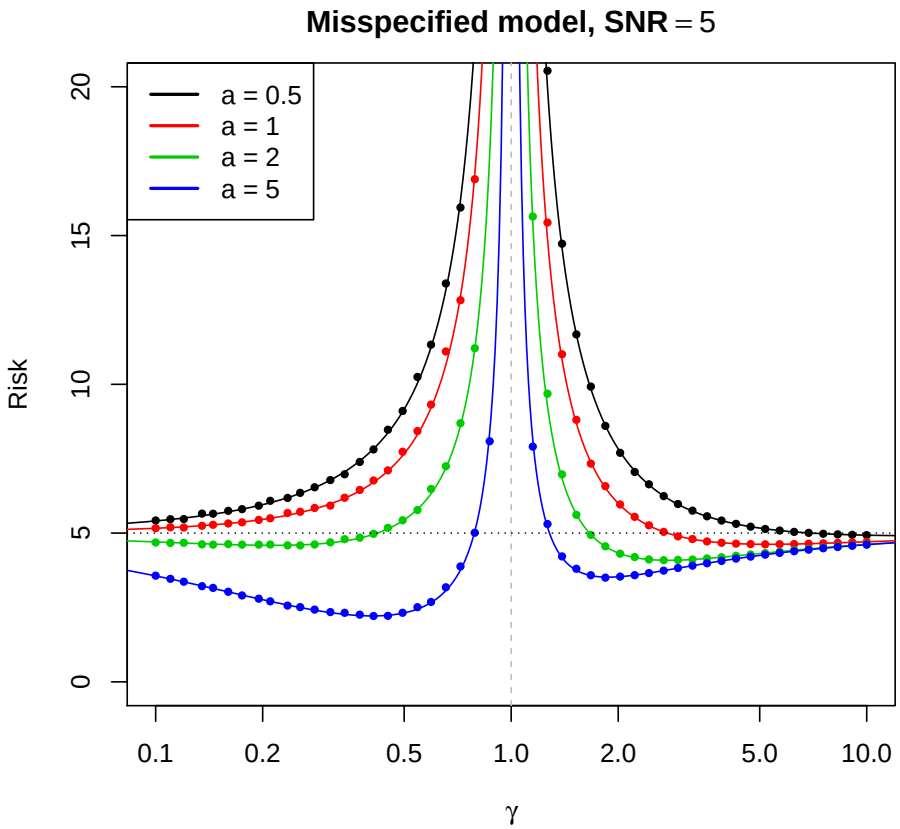
$$Y = \sum_{r=0}^{\infty} \beta_r X_r + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \tau^2), \quad \epsilon \perp\!\!\!\perp \vec{\boldsymbol{X}}_\infty, \quad \mathrm{V}(Y) < \infty.$$

$$X_0 = 1, \quad \{X_1, X_2, \ldots\} \text{ are jointly normal distributed,}$$

- The total prediction error under OLS has three parts:
  - τ^2: intrinsic error
  - $A(r) = \sum_{k=r+1}^{\infty} \Delta_k^2$: approx error, $\Delta_k^2 = V(Y|\vec{\boldsymbol{X}}_{r-1}) - V(Y|\vec{\boldsymbol{X}}_r)$
  - $\varepsilon(r,n) = \mathrm{E}_n\left[(\hat{\boldsymbol{\beta}}_r - \boldsymbol{\beta}_r^*)^\top \mathrm{E}(\vec{\boldsymbol{X}}_r \vec{\boldsymbol{X}}_r^\top)(\hat{\boldsymbol{\beta}}_r - \boldsymbol{\beta}_r^*)\right]$, where $\hat{\boldsymbol{\beta}}_r$ (or $\boldsymbol{\beta}_r^*$) is the training sample (or population) OLS coefficient.

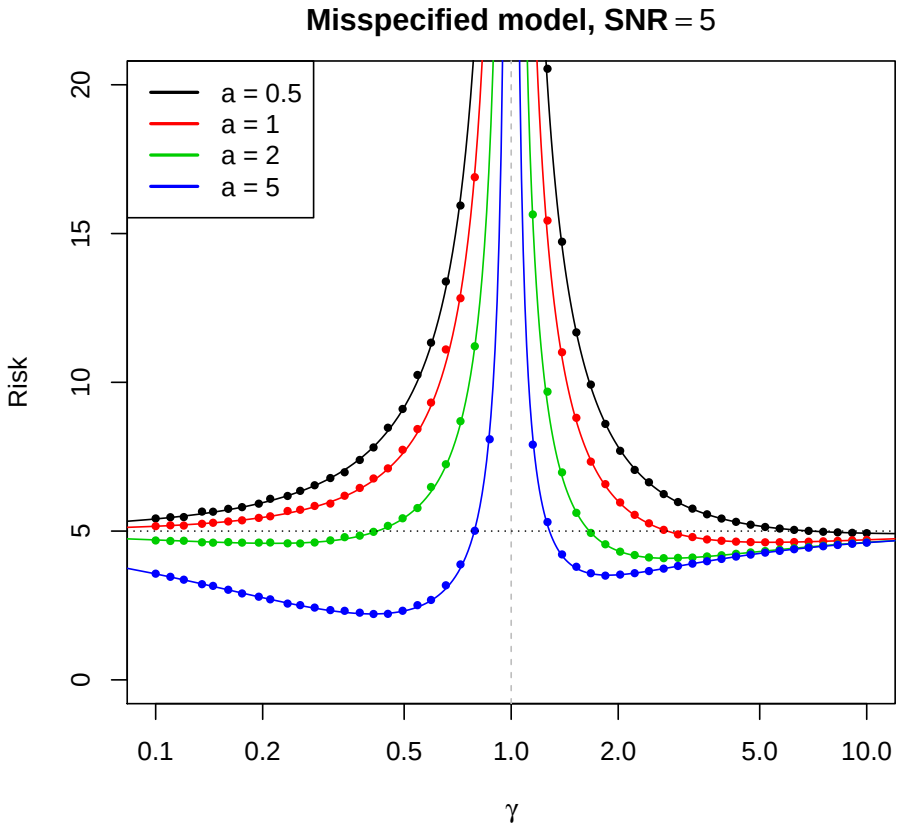
$$\varepsilon(r,n) = \frac{A(r) + \tau^2}{n - r - 2}\left(\frac{n-2}{n} + r\right)$$

Double Descent without over-fitting

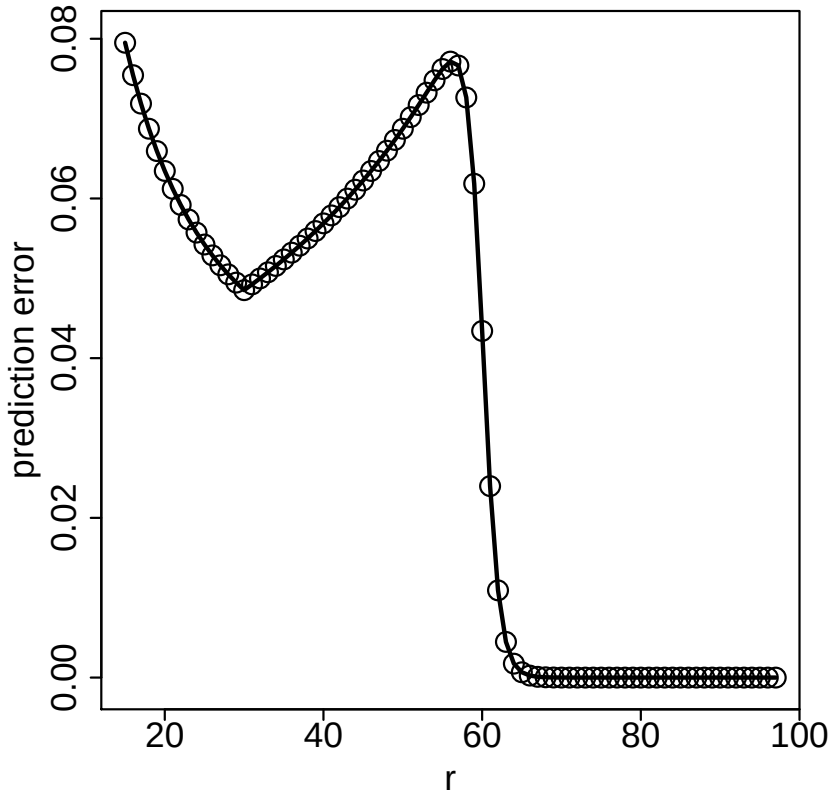


(a) $A(p)$ does not vary with p, $\gamma = p/n$ (**Hastie et al. (2019)**)

Double Descent without over-fitting

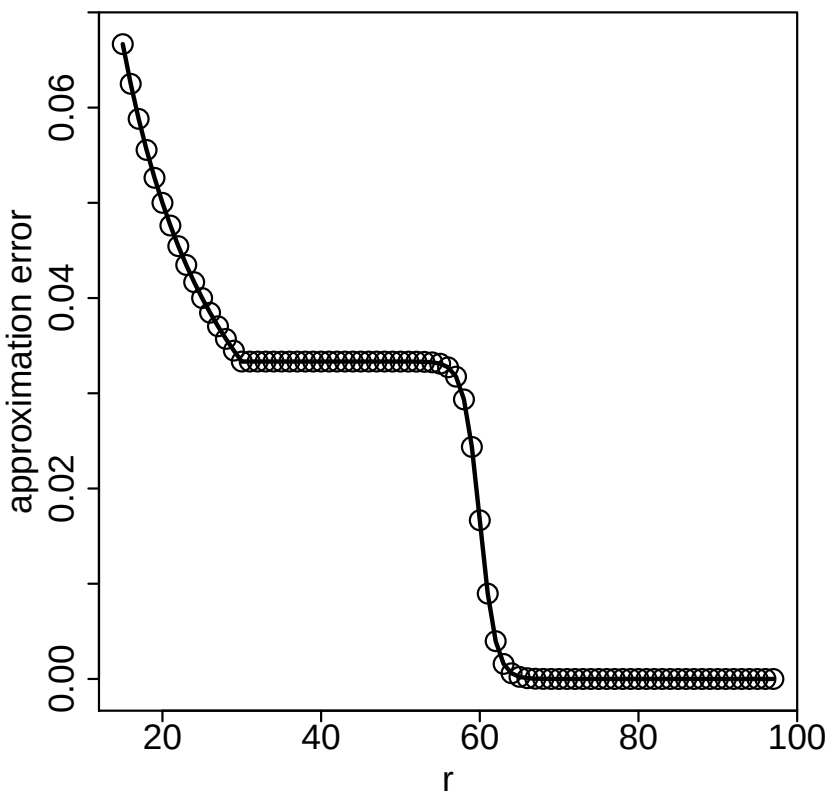


(a) $A(p)$ does not vary with p, $\gamma = p/n$ (**Hastie et al. (2019)**)

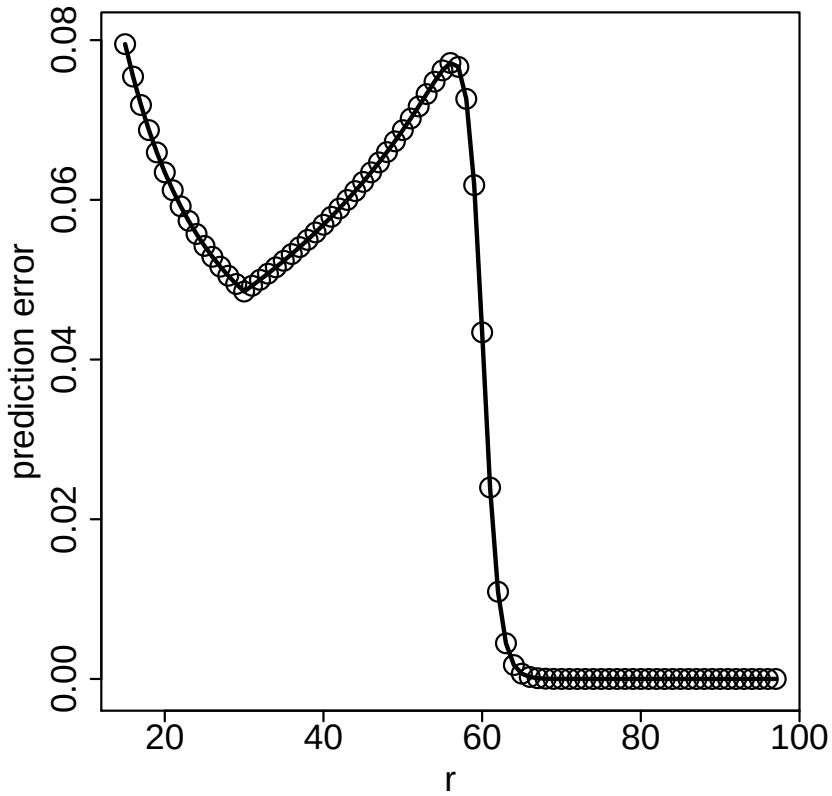


(b) $A(r)$ varies with r ($n = 100$)

Double Descent without over-fitting

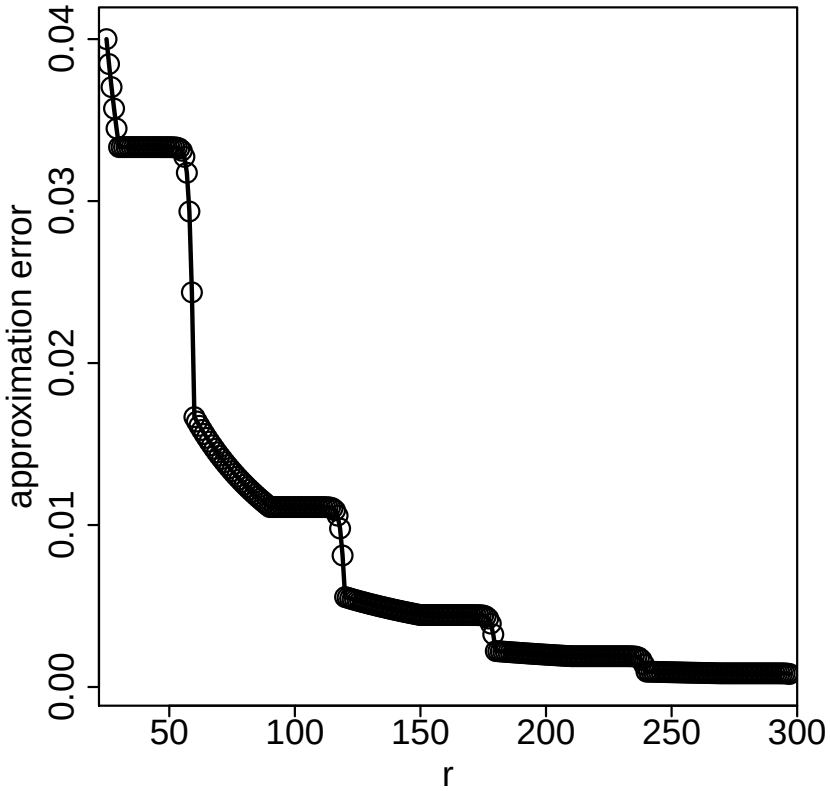


(a) pattern of approximation $A(r)$

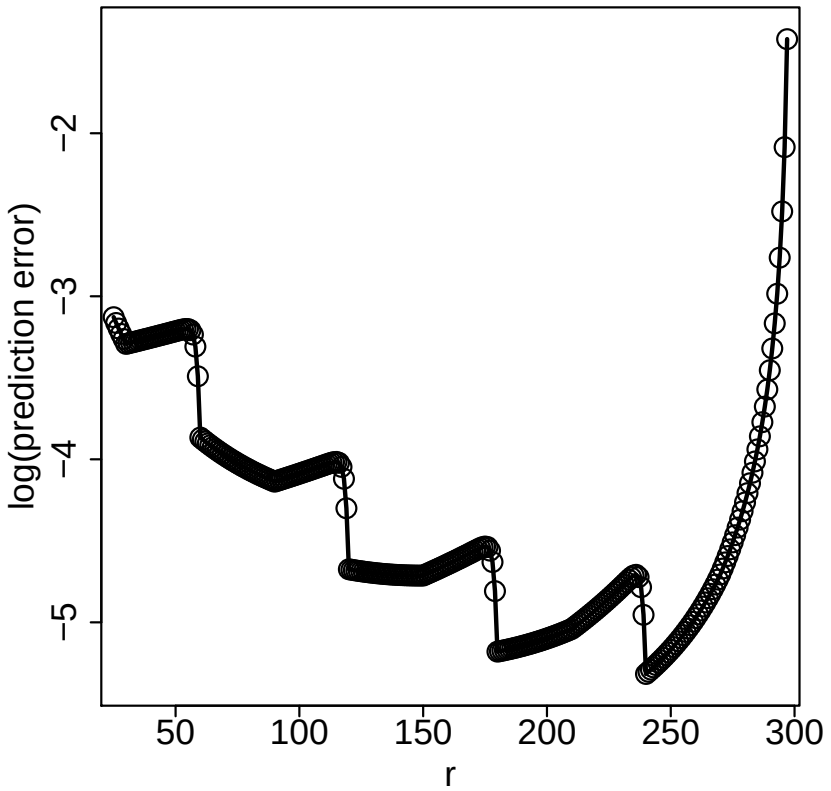


(b) prediction error at $n = 100$

Multiple Descents



(a) pattern of approximation $A(r)$



(b) prediction error at $n = 300$

Regression Tree with Potentially Infinite Depth

- Assume both target and training populations are the same, satisfying

$$X_1, X_2, \ldots \overset{i.i.d.}{\sim} \text{Bernoulli}(1/2), \quad \mathrm{V}(Y) < \infty,$$

and that dependence of Y on $\{X_1, X_2, \ldots\}$ is arbitrary.

Regression Tree with Potentially Infinite Depth

- Assume both target and training populations are the same, satisfying

$$X_1, X_2, \ldots \overset{i.i.d.}{\sim} \text{Bernoulli}(1/2), \quad \mathrm{V}(Y) < \infty,$$

and that dependence of Y on $\{X_1, X_2, \ldots\}$ is arbitrary.

- Prediction for a unit with covariates value $\vec{\boldsymbol{x}}_\infty$:

$$\hat{\boldsymbol{\mu}}_r(\vec{\boldsymbol{x}}_r) = \begin{cases} \{n(\vec{\boldsymbol{x}}_r)\}^{-1} \sum_{i:\vec{\boldsymbol{x}}_{ir}=\vec{\boldsymbol{x}}_r} Y_i, & \text{if } n(\vec{\boldsymbol{x}}_r) > 0, \\ \{n(\vec{\boldsymbol{x}}_k)\}^{-1} \sum_{i:\vec{\boldsymbol{x}}_{ik}=\vec{\boldsymbol{x}}_k} Y_i, & \text{if } n(\vec{\boldsymbol{x}}_k) > 0 \text{ and } n(\vec{\boldsymbol{x}}_{k+1}) = 0, \\ & \text{for some } 0 \le k < r. \end{cases}$$

Regression Tree with Potentially Infinite Depth

- Assume both target and training populations are the same, satisfying

$$X_1, X_2, \ldots \overset{i.i.d.}{\sim} \text{Bernoulli}(1/2), \quad \mathrm{V}(Y) < \infty,$$

and that dependence of Y on $\{X_1, X_2, \ldots\}$ is arbitrary.

- Prediction for a unit with covariates value $\vec{\boldsymbol{x}}_\infty$:

$$\hat{\boldsymbol{\mu}}_r(\vec{\boldsymbol{x}}_r) = \begin{cases} \{n(\vec{\boldsymbol{x}}_r)\}^{-1} \sum_{i:\vec{\boldsymbol{x}}_{ir}=\vec{\boldsymbol{x}}_r} Y_i, & \text{if } n(\vec{\boldsymbol{x}}_r) > 0, \\ \{n(\vec{\boldsymbol{x}}_k)\}^{-1} \sum_{i:\vec{\boldsymbol{x}}_{ik}=\vec{\boldsymbol{x}}_k} Y_i, & \text{if } n(\vec{\boldsymbol{x}}_k) > 0 \text{ and } n(\vec{\boldsymbol{x}}_{k+1}) = 0, \\ & \text{for some } 0 \le k < r. \end{cases}$$

- The above is the "highest resolution imputation", given the ordering of the covariates (which means there are other methods.)

Decomposition of the prediction error

- Model error, which is merely the intrinsic error:

$$\tau^2 \equiv \mathrm{E}[\mathrm{Var}(Y \mid \vec{\boldsymbol{X}}_\infty)]$$

Decomposition of the prediction error

- Model error, which is merely the intrinsic error:

$$\tau^2 \equiv \mathrm{E}[\mathrm{Var}(Y \mid \vec{\boldsymbol{X}}_\infty)]$$

- Approximation error:

$$A(r) = \sum_{k=r+1}^{\infty} [\mathrm{E}\{\mathrm{Var}(Y \mid \vec{\boldsymbol{X}}_{k-1})\} - \mathrm{E}\{\mathrm{Var}(Y \mid \vec{\boldsymbol{X}}_k)\}]$$

Decomposition of the prediction error

- Model error, which is merely the intrinsic error:

$$\tau^2 \equiv \mathrm{E}[\mathrm{Var}(Y \mid \vec{\boldsymbol{X}}_\infty)]$$

- Approximation error:

$$A(r) = \sum_{k=r+1}^{\infty} [\mathrm{E}\{\mathrm{Var}(Y \mid \vec{\boldsymbol{X}}_{k-1})\} - \mathrm{E}\{\mathrm{Var}(Y \mid \vec{\boldsymbol{X}}_k)\}]$$

- Estimation error: $\varepsilon(r,n) = \mathrm{E}_n\mathrm{E}[\{\hat{\boldsymbol{\mu}}_r(\vec{\boldsymbol{X}}_r) - \mathrm{E}(Y \mid \vec{\boldsymbol{X}}_r)\}^2]$

$$\varepsilon(r,n) = \left\{A(r) + \tau^2\right\} \cdot \mathrm{E}_n \left[\frac{\mathbb{1}(n(\vec{1}_r) > 0)}{n(\vec{1}_r)}\right]$$
$$+ \sum_{k=0}^{r-1} \left\{A(k) + \tau^2\right\} \cdot \mathrm{E}_n \left[\frac{\mathbb{1}(n(\vec{1}_k) > 0, n(\vec{1}_{k+1}) = 0)}{n(\vec{1}_k)}\right]$$
$$+ \sum_{k=0}^{r-1} \{A(k) - A(r)\} \cdot \mathrm{E}_n \left[\mathbb{1}(n(\vec{1}_k) > 0, n(\vec{1}_{k+1}) = 0)\right]$$

Stochastic World $\tau^2 > 0$: Optimal resolution R_n when $\mathcal{F}_r = \sigma\{X_1, \ldots, X_r\}$ and with n *i.i.d.* observations

- Resolution Loss: $A(r) = \sum_{i=r}^{\infty} \mathrm{E}[(\mu_{i+1} - \mu_i)^2 | \mathcal{F}_r]$
- Estimation loss: $\varepsilon(r, n) = \mathrm{E}(\mu_r - \hat{\mu}_r)^2$

Table: Optimal R_n and Minimal Loss $L_n \equiv \mathrm{PE}_n - \tau^2$ (as $n \to \infty$)

ε(r, n) \ A(r)	$e^{-\xi r}$	$r^{-\xi}$	$\log^{-\xi}(r)$
r^{α}/n $(\alpha \geq 1)$	$R_n = c_n \log n$	$c_n n^{1/(\alpha+\xi)}$	$\frac{c_n n^{1/\alpha}}{\log^{\xi/\alpha}(n)}$
	$L_n \propto \log^{\alpha}(n)/n$	$n^{-\xi/(\alpha+\xi)}$	$\log^{-\xi}(n)$
α^r/n $(\alpha > 1)$	$R_n = \frac{\log n + \log c_n}{\xi + \log \alpha}$	$c_n \log(n)$	$c_n \log(n)$
	$L_n \propto n^{-\xi/(\xi+\log \alpha)}$	$\log^{-\xi}(n)$	$\{\log\log(n)\}^{-\xi}$

- $c_n = O(1)$ but satisfies different constraints for different settings.

Deterministic World $\tau^2 = 0$, Optimal resolution R_n

Table: Optimal R_n and Minimal Loss $L_n \equiv \mathrm{PE}_n$ (as $n \to \infty$)

ε(r, n) \ A(r)	$e^{-\xi r}$	$r^{-\xi}$	$\log^{-\xi}(r)$
linear regression	$R_n = n - c_n$ $L_n \propto n e^{-\xi n}$	$c_n n$ $n^{-\xi}$	$c_n n^k,\ k \in (0, 1]$ $\log^{-\xi}(n)$
regression	$R_n \begin{cases} \gg \text{or} = c_n \log(n) \\ = c_n \log(n) \\ = c_n \log(n) \end{cases}$	$c_n \log(n)$	$c_n \log(n)$
tree	$L_n \propto \begin{cases} n^{-1}, & \xi > \log(2) \\ \log(n)/n, & \xi = \log(2) \\ n^{-\xi/\log(2)}, & \xi < \log(2) \end{cases}$	$\log^{-\xi}(n)$	$\{\log\log(n)\}^{-\xi}$

- $c_n = O(1)$ but satisfies different constraints for different settings.

What about the ordering of covariates?

- Suppose a second ordering $\{k_0, k_1, \ldots, k_r\}$ includes the *first* $r - M_r$ predictors of the first ordering; i.e., M_r is the minimum integer s.t. $\{0, 1, 2, \ldots, r - M_r\} \subset \{k_0, \ldots, k_r\}$

What about the ordering of covariates?

- Suppose a second ordering $\{k_0, k_1, \ldots, k_r\}$ includes the *first* $r - M_r$ predictors of the first ordering; i.e., M_r is the minimum integer s.t. $\{0, 1, 2, \ldots, r - M_r\} \subset \{k_0, \ldots, k_r\}$

Then the optimal rate under the second ordering is no worse than that under the first ordering if

- Exponential decaying: $\limsup M_r \leq C$

What about the ordering of covariates?

- Suppose a second ordering $\{k_0, k_1, \ldots, k_r\}$ includes the *first* $r - M_r$ predictors of the first ordering; i.e., M_r is the minimum integer s.t. $\{0, 1, 2, \ldots, r - M_r\} \subset \{k_0, \ldots, k_r\}$

Then the optimal rate under the second ordering is no worse than that under the first ordering if

- Exponential decaying: $\limsup M_r \leq C$
- Polynomial decaying: $\limsup M_r / r < 1$

What about the ordering of covariates?

- Suppose a second ordering $\{k_0, k_1, \ldots, k_r\}$ includes the *first* $r - M_r$ predictors of the first ordering; i.e., M_r is the minimum integer s.t. $\{0, 1, 2, \ldots, r - M_r\} \subset \{k_0, \ldots, k_r\}$

Then the optimal rate under the second ordering is no worse than that under the first ordering if

- Exponential decaying: $\limsup M_r \leq C$
- Polynomial decaying: $\limsup M_r/r < 1$
- Logarithm decaying: $M_r = r - r^{1/a_r}$, where $a_r = O(1)$.

What about the ordering of covariates?

- Suppose a second ordering $\{k_0, k_1, \ldots, k_r\}$ includes the *first* $r - M_r$ predictors of the first ordering; i.e., M_r is the minimum integer s.t. $\{0, 1, 2, \ldots, r - M_r\} \subset \{k_0, \ldots, k_r\}$

Then the optimal rate under the second ordering is no worse than that under the first ordering if

- Exponential decaying: $\limsup M_r \leq C$
- Polynomial decaying: $\limsup M_r/r < 1$
- Logarithm decaying: $M_r = r - r^{1/a_r}$, where $a_r = O(1)$.

What about the ordering of covariates?

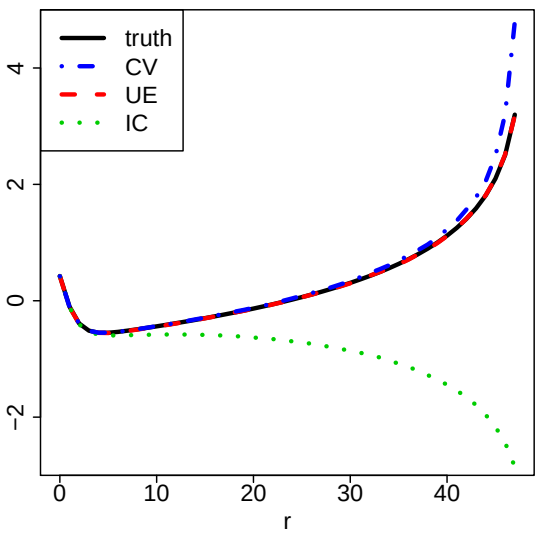
- Suppose a second ordering $\{k_0, k_1, \ldots, k_r\}$ includes the *first* $r - M_r$ predictors of the first ordering; i.e., M_r is the minimum integer s.t. $\{0, 1, 2, \ldots, r - M_r\} \subset \{k_0, \ldots, k_r\}$

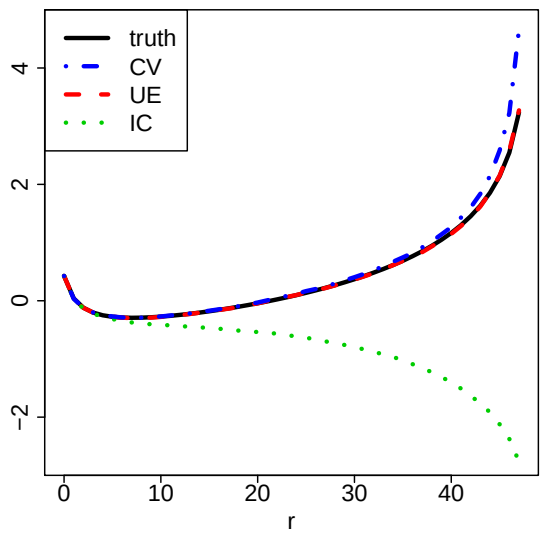
Then the optimal rate under the second ordering is no worse than that under the first ordering if

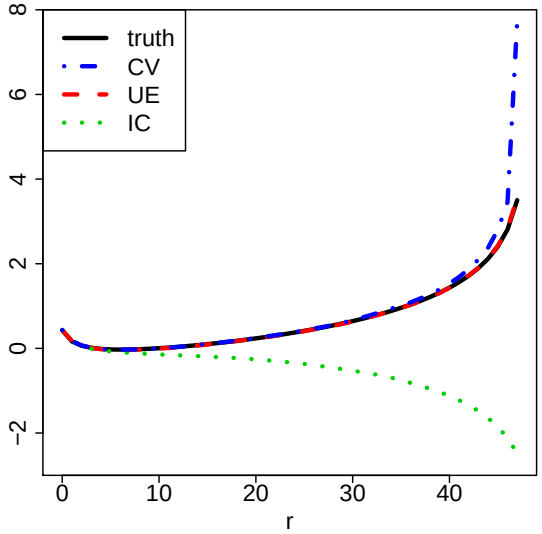
- Exponential decaying: $\limsup M_r \leq C$
- Polynomial decaying: $\limsup M_r/r < 1$
- Logarithm decaying: $M_r = r - r^{1/a_r}$, where $a_r = O(1)$.

- The slower $A(r)$ decays, the less restriction on $\{M_r\}$.

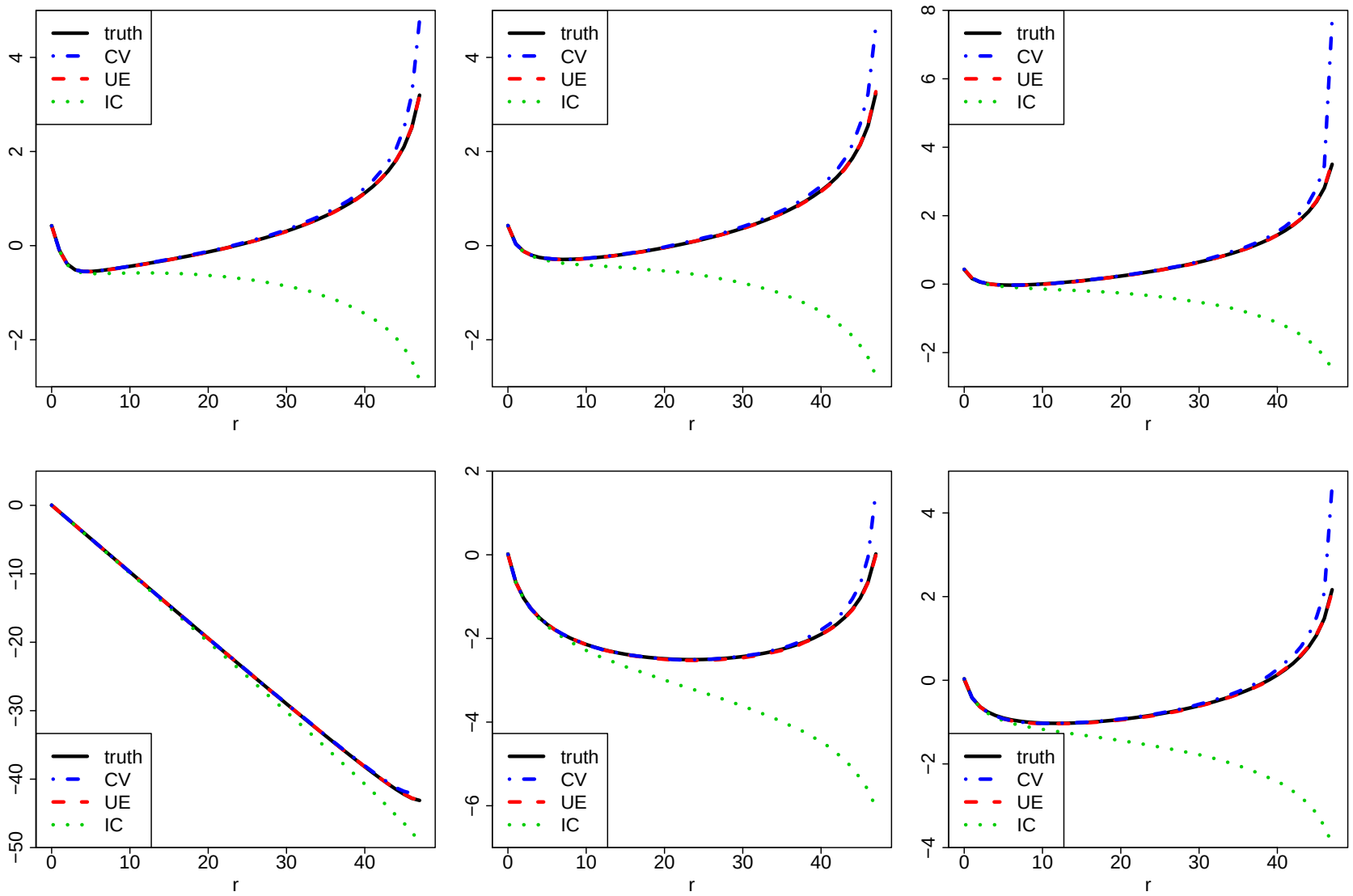
Numerical Illustrations Under Linear Models







Numerical Illustrations Under Linear Models



(a) Exponential (b) Polynomial (c) Logarithmic

Performance of estimators when $\tau^2 > 0$

Table: Estimated resolution and corresponding prediction error when $n = 50$.

Type / r_{opt} / $\text{PE}_n(r_{\text{opt}})$	Method	$\hat{R}$	95% QR	$\text{PE}_n(\hat{R})/\text{PE}_n(r_{\text{opt}})$	95% QR
Exponential	Oracle	–	–	1.00	[0.92, 1.18]
$r_{\text{opt}} = 4$	CV	6	[2, 20]	1.34	[0.92, 1.72]
$\text{PE}_n(r_{\text{opt}}) = 0.5767$	UE	7	[2, 47]	1.93	[0.92, 9.23]
	IC	47	[46, 47]	19.96	[3.86, 81.00]
Polynomial	Oracle	–	–	1.00	[0.87, 1.24]
$r_{\text{opt}} = 7$	CV	10	[2, 44]	1.37	[0.91, 5.17]
$\text{PE}_n(r_{\text{opt}}) = 0.7463$	UE	11	[2, 47]	1.62	[0.90, 8.18]
	IC	47	[45, 47]	14.77	[2.93, 57.54]
Logarithmic	Oracle	–	–	1.00	[0.89, 1.21]
$r_{\text{opt}} = 6$	CV	9	[2, 41]	1.38	[0.91, 3.97]
$\text{PE}_n(r_{\text{opt}}) = 0.9714$	UE	10	[2, 46]	1.87	[0.91, 8.04]
	IC	47	[46, 47]	14.60	[3.52, 58.19]

Performance of estimators when $\tau^2 = 0$

Table: Estimated resolution and corresponding prediction error when $n = 50$.

Type / r_{opt} / $\text{PE}_n(r_{\text{opt}})$	Method	$\hat{R}$	95% QR	$\text{PE}_n(\hat{R})/\text{PE}_n(r_{\text{opt}})$	95% QR
Exponential	Oracle	–	–	1.08	[0.12, 4.38]
$r_{\text{opt}} = 47$	CV	46	[44, 47]	3.57	[0.33, 15.54]
$\text{PE}_n(r_{\text{opt}}) = 1.90 \times 10^{-19}$	UE	46	[45, 47]	2.30	[0.32, 12.34]
	IC	47	[47, 47]	1.40	[0.29, 6.68]
Polynomial	Oracle	–	–	0.99	[0.71, 1.48]
$r_{\text{opt}} = 23$	CV	27	[13, 47]	1.31	[0.73, 4.02]
$\text{PE}_n(r_{\text{opt}}) = 0.0816$	UE	28	[13, 47]	1.44	[0.74, 4.68]
	IC	47	[46, 47]	5.69	[1.28, 23.97]
Logarithmic	Oracle	–	–	1.00	[0.83, 1.34]
$r_{\text{opt}} = 12$	CV	15	[4, 44]	1.38	[0.87, 5.39]
$\text{PE}_n(r_{\text{opt}}) = 0.357$	UE	16	[5, 47]	1.70	[0.87, 7.72]
	IC	47	[46, 47]	10.99	[2.27, 41.35]

For those who just woke up ...

Big data for Individualized Inference/Prediction

- Statistics becomes an approximation scheme: **approximating individuals by proxy populations**, not the other way around.

For those who just woke up ...

Big data for Individualized Inference/Prediction

- Statistics becomes an approximation scheme: **approximating individuals by proxy populations**, not the other way around.
- **The concept of bias is more critical than (pure) variance**, which exists essentially only at the infinite resolution level.

For those who just woke up ...

Big data for Individualized Inference/Prediction

- Statistics becomes an approximation scheme: **approximating individuals by proxy populations**, not the other way around.
- **The concept of bias is more critical than (pure) variance**, which exists essentially only at the infinite resolution level.
- In the stochastic world, a central task is to determine the **appropriate level of approximation via the bias-variance trade-off.**

For those who just woke up ...

Big data for Individualized Inference/Prediction

- Statistics becomes an approximation scheme: **approximating individuals by proxy populations**, not the other way around.
- **The concept of bias is more critical than (pure) variance**, which exists essentially only at the infinite resolution level.
- In the stochastic world, a central task is to determine the **appropriate level of approximation via the bias-variance trade-off.**
- **In the deterministic world, we can put all our eggs in the basket of bias**, when we have great sparsity.

For those who just woke up ...

Big data for Individualized Inference/Prediction

- Statistics becomes an approximation scheme: **approximating individuals by proxy populations**, not the other way around.
- **The concept of bias is more critical than (pure) variance**, which exists essentially only at the infinite resolution level.
- In the stochastic world, a central task is to determine the **appropriate level of approximation via the bias-variance trade-off.**
- **In the deterministic world, we can put all our eggs in the basket of bias**, when we have great sparsity.

For those who just woke up ...

Big data for Individualized Inference/Prediction

- Statistics becomes an approximation scheme: **approximating individuals by proxy populations**, not the other way around.
- **The concept of bias is more critical than (pure) variance**, which exists essentially only at the infinite resolution level.
- In the stochastic world, a central task is to determine the **appropriate level of approximation via the bias-variance trade-off.**
- **In the deterministic world, we can put all our eggs in the basket of bias**, when we have great sparsity.

Resolution is a fundamental concept for scientific inference

- Low resolutions render operational meaning to (frequentest) probability and scientific evaluations

For those who just woke up ...

Big data for Individualized Inference/Prediction

- Statistics becomes an approximation scheme: **approximating individuals by proxy populations**, not the other way around.
- **The concept of bias is more critical than (pure) variance**, which exists essentially only at the infinite resolution level.
- In the stochastic world, a central task is to determine the **appropriate level of approximation via the bias-variance trade-off.**
- **In the deterministic world, we can put all our eggs in the basket of bias**, when we have great sparsity.

Resolution is a fundamental concept for scientific inference

- Low resolutions render operational meaning to (frequentest) probability and scientific evaluations
- Data, inference, decisions, evaluations may all have different resolutions (Meng, 2021, "Part II")

Detailed Result

Searching for optimal R_n in the Stochastic World: $\tau^2 > 0$

- $\frac{r_n}{n} = o(1)$ is necessary for $\varepsilon(r_n, n) = o(1)$, under which $\varepsilon(r_n, n) \asymp \frac{r_n}{n}$.
- **General Result:** Let R_n be a rate-optimal resolution, and $L_n = A(R_n) + \varepsilon(R_n, n)$ be the minimal prediction error (excluding τ^2). Assume *polynomial estimation error*, that is, $\varepsilon(r, n) \asymp r^\alpha / n$.
  - (i) Hard Thresholding: $A(r) = 0$ for $r \geq r_0$, and $A(r) > 0$ for $r < r_0$. Then $R_n \asymp$ *Constant* with the constraint that $\liminf_{n\to\infty} R_n \geq r_0$; and $L_n \asymp n^{-1}$.
  - (ii) Exponential $A(r) \asymp e^{-\xi r}$. Then $R_n = a_n \log(n)$ with a_n satisfying $a_n \asymp 1$ and $n^{1-\xi a_n} \log^{-\alpha}(n) = O(1)$; and $L_n \asymp n^{-1} \log^{\alpha}(n)$.
  - (iii) Polynomial $A(r) \asymp r^{-\xi}$. Then $R_n \asymp n^{1/(\alpha+\xi)}$; and $L_n \asymp n^{-\xi/(\alpha+\xi)}$.
  - (iv) Logarithmic $A(r) \asymp \log^{-\xi}(r)$. Then $R_n = a_n n^{1/\alpha} \log^{-\xi/\alpha}(n)$ with a_n satisfying $a_n = O(1)$ and $\limsup_{n\to\infty} \left[\log^{-1}(n) \log(a_n^{-1})\right] < \alpha^{-1}$; and $L_n \asymp \log^{-\xi}(n)$.

Search for Optimal R_n in the Deterministic World: $\tau^2 = 0$

- The prediction error simplifies to $A(r) \cdot \frac{(n+1)(n-2)}{n(n-r-2)}$.

  Specific Results (for normal model)

  (i) Hard Thresholding: $A(r) = 0$ for $r \geq r_0$, and $A(r) > 0$ for $r < r_0$. The optimal resolution is any (sequence) R_n such that $\liminf_{n\to\infty} R_n \geq r_0$ and $R_n \leq n - 3$; and $L_n = 0$.

  (ii) Exponential $A(r) \asymp e^{-\xi r}$. $R_n = n - O(1)$ with $R_n \leq n - 3$; and $L_n \asymp ne^{-\xi n}$.

  (iii) Polynomial $A(r) \asymp r^{-\xi}$. $R_n = a_n n$ with a_n satisfying $a_n \asymp 1$ and $\limsup a_n < 1$; and $L_n \asymp n^{-\xi}$.

  (iv) Logarithmic $A(r) \asymp \log^{-\xi}(r)$. Optimal resolution is any (sequence) R_n such that $\limsup R_n/n < 1$, $\liminf \frac{\log R_n}{\log n} > 0$; and $L_n \asymp \log^{-\xi}(n)$.

Search for Optimal R_n in the Stochastic World: $\tau^2 > 0$

- $\frac{2^r}{n} = o(1)$ is necessary for $\varepsilon(r,n) = o(1)$, under which $\varepsilon(r,n) \asymp \frac{2^r}{n}$.
- **General Result**: Assume *exponential estimation error*: $\varepsilon(r,n) \asymp \frac{\alpha^r}{n}$.
  - (i) Hard Thresholding: $A(r) = 0$ for $r \geq r_0$, and $A(r) > 0$ for $r < r_0$. Then $R_n \asymp$ *Constant* with the constraint that $\liminf_{n\to\infty} R_n \geq r_0$, and $L_n \asymp n^{-1}$.
  - (ii) Exponential $A(r) \asymp e^{-\xi r}$. Then $R_n = \frac{\log(n)+\log(a_n)}{\log(\alpha)+\xi}$ with $a_n \asymp 1$; and $L_n \asymp n^{-\xi/\{\log(\alpha)+\xi\}}$.
  - (iii) Polynomial $A(r) \asymp r^{-\xi}$. Then $R_n = a_n \log(n)$ with a_n satisfying $a_n \asymp 1$ and $n^{a_n \log(\alpha)-1}\log^{\xi}(n) = O(1)$; and $L_n \asymp \log^{-\xi}(n)$.
  - (iv) Logarithmic $A(r) \asymp \log^{-\xi}(r)$. Then $R_n = a_n \log(n)$ with a_n satisfying

$$\liminf_{n\to\infty} \frac{\log(a_n)}{\log\log(n)} > -1, \quad \text{and} \quad \frac{\{\log\log(n)\}^{\xi}}{n^{1-a_n\log(\alpha)}} = O(1);$$

and $L_n \asymp \{\log\log(n)\}^{-\xi}$.

Search optimal R_n in the deterministic World: $\sigma^2 = 0$

Let $L_n = A(R_n) + \varepsilon(R_n, n) \le A(R_n) + \bar{\varepsilon}(R_n, n) \equiv \overline{L}_n$. **Specific Results:**

(i) Hard Thresholding: $A(r) = 0$ for $r \ge r_0$, and $A(r) > 0$ for $r < r_0$.

Then R_n satisfies that $\liminf_{n\to\infty} R_n \ge r_0$; and $L_n \asymp (1 - 2^{-r_0})^n$.

(ii) Exponential $A(r) \asymp e^{-\xi r}$.

(a) If $e^{-\xi} < 1/2$, then $\overline{R}_n$ satisfies $ne^{-\xi R_n} = O(1)$; and $\overline{L}_n \asymp n^{-1}$.

(b) If $e^{-\xi} = 1/2$, then $\overline{R}_n = a_n \log(n)$ with a_n satisfying $a_n \asymp 1$ and $n^{1-a_n \log(2)}/\log(n) = O(1)$; and $\overline{L}_n \asymp n^{-1}\log(n)$.

(c) If $e^{-\xi} > 1/2$, then $\overline{R}_n = a_n \log(n)$ with a_n satisfying $n^{a_n - 1/\log(2)} \asymp 1$; and $\overline{L}_n \asymp n^{-\xi/\log(2)}$.

(iii) Polynomial $A(r) \asymp r^{-\xi}$. Then $\overline{R}_n = a_n \log(n)$ with a_n satisfying $a_n \asymp 1$ and $n^{a_n \log(2) - 1} = O(1)$; and $\overline{L}_n \asymp \log^{-\xi}(n)$.

(iv) Logarithmic $A(r) \asymp \log^{-\xi}(r)$. Then $\overline{R}_n = a_n \log(n)$ with a_n satisfying

$$\liminf_{n\to\infty} \frac{\log(a_n)}{\log\log(n)} > -1, \quad \text{and} \quad n^{a_n \log(2) - 1} = O(1);$$

and $\overline{L}_n \asymp \{\log\log(n)\}^{-\xi}$.

The Deterministic World without Variance

An asymptotic lower bound for the estimation error $\varepsilon(r, n)$ has the following form.

(i) Hard Thresholding: $A(r) = 0$ for $r \geq r_0$, and $A(r) > 0$ for $r < r_0$. Then $A(r_n) + \varepsilon(r, n) \gtrsim (1 - 2^{-r_0})^n$.

(ii) Exponential $A(r) \asymp e^{-\xi r}$. Then $A(r_n) + \varepsilon(r, n) \gtrsim n^{-\xi/\log(2)}$.

(iii) Polynomial $A(r) \asymp r^{-\xi}$. Then $A(r_n) + \varepsilon(r, n) \gtrsim \log^{-\xi}(n)$.

(iv) Logarithmic $A(r) \asymp \log^{-\xi}(r)$. Then $A(r_n) + \varepsilon(r, n) \gtrsim \{\log\log(n)\}^{-\xi}$.