High Dimensional M-Estimation & Inference from Observational Data with Incomplete Responses

A Semi-Parametric Doubly Robust Framework

Abhishek Chakrabortty[1]

Department of Statistics
University of Pennsylvania

Group Meeting
April 24, 2019

[1] Joint work with Jiarui Lu, T. Tony Cai and Hongzhe Li.

Big Data Era: The Challenges of Incomplete Information

- Current era of 'big data' and data science $\rightsquigarrow$ rapid influx of large **and** high dimensional data (easily available and computationally tractable).

- Rich information on multitudes of variables at the same place $\rightsquigarrow$ many interesting scientific questions and also unique statistical challenges!

- Current era of 'big data' and data science $\rightsquigarrow$ rapid influx of large **and** high dimensional data (easily available and computationally tractable).
- Rich information on multitudes of variables at the same place $\rightsquigarrow$ many interesting scientific questions and also unique statistical challenges!
- One frequently encountered challenge: incompleteness of the data and in particular, (partial) missingness of the response of interest.
  - Reasons could be 'circumstantial' (e.g. practical constraints such as logistics, time, cost issues etc.), or it could be 'by design' (e.g. due to the 'treatment' assignment/non-assignment mechanism).
  - The response corresponding to a 'treatment' of interest could not be observed for a person who is not 'treated' (and vice versa).

Big Data Era: The Challenges of Incomplete Information

- Current era of 'big data' and data science $\rightsquigarrow$ rapid influx of large **and** high dimensional data (easily available and computationally tractable).
- Rich information on multitudes of variables at the same place $\rightsquigarrow$ many interesting scientific questions and also unique statistical challenges!
- One frequently encountered challenge: incompleteness of the data and in particular, (partial) missingness of the response of interest.
  - Reasons could be 'circumstantial' (e.g. practical constraints such as logistics, time, cost issues etc.), or it could be 'by design' (e.g. due to the 'treatment' assignment/non-assignment mechanism).
  - The response corresponding to a 'treatment' of interest could not be observed for a person who is not 'treated' (and vice versa).
- Another complication in both cases: observational nature of the data. The missingness mechanism could be informative (not randomized)!

Challenges of Incompleteness Contd. and Relevance in Modern Studies

- Observational data $\rightsquigarrow$ typically informative missingness (or treatment assignment) mechanism. Could **depend** on the person's covariates.
- Often termed selection bias or treatment by indication or confounding (in causal inference) in observational studies. **Has** to be factored in!
- Need to account for the missingness in a proper principled way under minimal conditions to ensure valid, unbiased (and robust) inference.

Challenges of Incompleteness Contd. and Relevance in Modern Studies

- Observational data $\rightsquigarrow$ typically informative missingness (or treatment assignment) mechanism. Could **depend** on the person's covariates.
- Often termed selection bias or treatment by indication or confounding (in causal inference) in observational studies. **Has** to be factored in!
- Need to account for the missingness in a proper principled way under minimal conditions to ensure valid, unbiased (and robust) inference.
- Relevance: these issues occur in virtually **any** modern day large scale observational study arising in various scientific disciplines, including:
- Biomedical studies (e.g. electronic health records (EHR) data); and Integrative genomics (e.g. gene expression data and eQTL studies).
- Also econometrics (policy evaluation), computer science, finance etc.

- **Variables of interest:** outcome $Y \in \mathcal{Y} \subseteq \mathbb{R}$ and covariates $\mathbf{X} \in \mathcal{X} \subseteq \mathbb{R}^p$ (possibly high dimensional, compared to the sample size).
  - The supports $\mathcal{Y}$ and $\mathcal{X}$ of Y and $\mathbf{X}$ need **not** be continuous.
- **Main issue:** Y may not always be observed. Let $T \in \{0, 1\}$ denote the indicator of the true Y being observed.
- The (partly) unobserved random vector $(T, Y, \mathbf{X})$ is assumed to be jointly defined on a common probability space with measure $\mathbb{P}(\cdot)$.

- **Variables of interest:** outcome $Y \in \mathcal{Y} \subseteq \mathbb{R}$ and covariates $\mathbf{X} \in \mathcal{X} \subseteq \mathbb{R}^p$ (possibly high dimensional, compared to the sample size).
    - The supports $\mathcal{Y}$ and $\mathcal{X}$ of Y and $\mathbf{X}$ need **not** be continuous.
- **Main issue:** Y may not always be observed. Let $T \in \{0, 1\}$ denote the indicator of the true Y being observed.
- The (partly) unobserved random vector $(T, Y, \mathbf{X})$ is assumed to be jointly defined on a common probability space with measure $\mathbb{P}(\cdot)$.
- Observable data: $\mathcal{D}_n := \{\mathbf{Z}_i := (T_i, T_i Y_i, \mathbf{X}_i) : i = 1, \ldots, n\} \overset{iid}{\sim} \mathbf{Z}$, where $\mathbf{Z} := (T, TY, \mathbf{X})$ whose distribution is defined via $\mathbb{P}(\cdot)$.
- **High dimensional setting:** p can diverge with n (including $p \gg n$).

Applicability of the Framework

- Generally applicable to **any** missing data setting - with missing outcomes Y and (possibly) high dimensional covariates $\mathbf{X}$.
- Causal inference problems (via 'potential' outcomes framework).

- Generally applicable to **any** missing data setting - with missing outcomes Y and (possibly) high dimensional covariates $\mathbf{X}$.
- Causal inference problems (via 'potential' outcomes framework).
  - Here, $\mathbf{X}$ is often called 'confounders' (for observational studies) or 'adjustment' variables/features (for randomized trials).
  - **Usual set-up:** binary 'treatment' (a.k.a. exposure/intervention) assignment: $\mathcal{T} \in \{0, 1\}$, and potential outcomes: $\{Y_{(0)}, Y_{(1)}\}$.
  - **Observed outcome:** $\mathbb{Y} := Y_{(0)}1(\mathcal{T} = 0) + Y_{(1)}1(\mathcal{T} = 1)$, i.e. depending on $\mathcal{T}$, we observe only **one** of $\{Y_{(0)}, Y_{(1)}\}$.

- Generally applicable to **any** missing data setting - with missing outcomes Y and (possibly) high dimensional covariates $\mathbf{X}$.

- Causal inference problems (via 'potential' outcomes framework).
  - Here, $\mathbf{X}$ is often called 'confounders' (for observational studies) or 'adjustment' variables/features (for randomized trials).
  - **Usual set-up:** binary 'treatment' (a.k.a. exposure/intervention) assignment: $\mathcal{T} \in \{0, 1\}$, and potential outcomes: $\{Y_{(0)}, Y_{(1)}\}$.
  - **Observed outcome:** $\mathbb{Y} := Y_{(0)} 1(\mathcal{T} = 0) + Y_{(1)} 1(\mathcal{T} = 1)$, i.e. depending on $\mathcal{T}$, we observe only **one** of $\{Y_{(0)}, Y_{(1)}\}$.
  - For **each** $j \in \{0, 1\}$, this set-up is **included** based on the 'map':

$$(T, Y, \mathbf{X}) \leftarrow (\mathcal{T}_j, Y_{(j)}, \mathbf{X}) \text{ with } \mathcal{T}_j := 1(\mathcal{T} = j) \;\; \forall\, j \in \{0, 1\}.$$

Applicability of the Framework

- Generally applicable to **any** missing data setting - with missing outcomes Y and (possibly) high dimensional covariates $\mathbf{X}$.

- Causal inference problems (via 'potential' outcomes framework).

  - Here, $\mathbf{X}$ is often called 'confounders' (for observational studies) or 'adjustment' variables/features (for randomized trials).

  - **Usual set-up:** binary 'treatment' (a.k.a. exposure/intervention) assignment: $\mathcal{T} \in \{0, 1\}$, and potential outcomes: $\{Y_{(0)}, Y_{(1)}\}$.

  - **Observed outcome:** $\mathbb{Y} := Y_{(0)}1(\mathcal{T} = 0) + Y_{(1)}1(\mathcal{T} = 1)$, i.e. depending on $\mathcal{T}$, we observe only **one** of $\{Y_{(0)}, Y_{(1)}\}$.

  - For **each** $j \in \{0, 1\}$, this set-up is **included** based on the 'map':

$$(T, Y, \mathbf{X}) \leftarrow (\mathcal{T}_j, Y_{(j)}, \mathbf{X}) \;\; \text{with } \mathcal{T}_j := 1(\mathcal{T} = j) \;\; \forall\, j \in \{0, 1\}.$$

  The case of any multi-category treatment also similarly included.

1. **Ignorability assumption:** $T \perp\!\!\!\perp Y \mid \mathbf{X}$.
   - A.k.a. 'missing at random' (MAR) in the missing data literature.
   - A.k.a. 'no unmeasured confounding' (NUC) in causal inference.
   - Special case: $T \perp\!\!\!\perp (Y, \mathbf{X})$. A.k.a. missing completely at random (MCAR) in missing data literature, and complete randomization (e.g. randomized trials) in causal inference (CI) literature.

1. **Ignorability assumption:** $T \perp\!\!\!\perp Y \mid \mathbf{X}$.
   - A.k.a. 'missing at random' (MAR) in the missing data literature.
   - A.k.a. 'no unmeasured confounding' (NUC) in causal inference.
   - Special case: $T \perp\!\!\!\perp (Y, \mathbf{X})$. A.k.a. missing completely at random (MCAR) in missing data literature, and complete randomization (e.g. randomized trials) in causal inference (CI) literature.

2. **Positivity assumption** (a.k.a. 'sufficient overlap' in CI literature):
   - Let $\pi(\mathbf{X}) := \mathbb{P}(T = 1 \mid \mathbf{X})$ be the propensity score (PS), and let $\pi_0 := \mathbb{P}(T = 1)$. Then, $\pi(\cdot)$ is uniformly bounded away from 0:

$$1 \geq \pi(\mathbf{x}) \geq \delta_\pi > 0 \quad \forall\, \mathbf{x} \in \mathcal{X}, \text{ for some constant } \delta_\pi > 0.$$

- **Rich resources** of data for discovery research; fast growing literature.

nature REVIEWS GENETICS

Review Article | Published: 18 May 2011

Using electronic health records to drive discovery in disease genomics

Isaac S. Kohane

Nature Reviews Genetics **12**, 417–428 (2011)

nature REVIEWS GENETICS

Review Article | Published: 02 May 2012

Mining electronic health records: towards better research applications and clinical care

Peter B. Jensen, Lars J. Jensen & Søren Brunak

Nature Reviews Genetics **13**, 395–405 (2012)

- **Rich resources** of data for discovery research; fast growing literature.

nature REVIEWS GENETICS

Review Article | Published: 18 May 2011

Using electronic health records to drive discovery in disease genomics

Isaac S. Kohane

Nature Reviews Genetics **12**, 417–428 (2011)

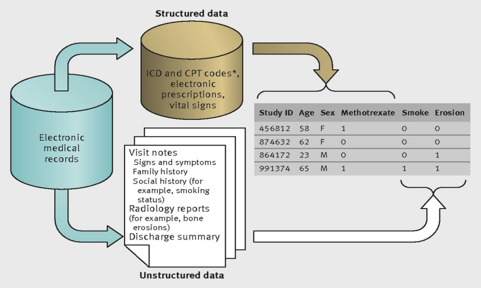
nature REVIEWS GENETICS

Review Article | Published: 02 May 2012

Mining electronic health records: towards better research applications and clinical care

Peter B. Jensen, Lars J. Jensen & Søren Brunak

Nature Reviews Genetics **13**, 395–405 (2012)



Study ID	Age	Sex	Methotrexate	Smoke	Erosion
456812	58	F	1	0	0
874632	62	F	0	0	0
864172	23	M	0	0	1
991374	65	M	1	1	1

- **Detailed** clinical and phenotypic data collected electronically for large patient cohorts, as part of routine health care delivery.

- **Rich resources** of data for discovery research; fast growing literature.


nature REVIEWS GENETICS

Review Article | Published: 18 May 2011

Using electronic health records to drive discovery in disease genomics

Isaac S. Kohane

Nature Reviews Genetics **12**, 417–428 (2011)

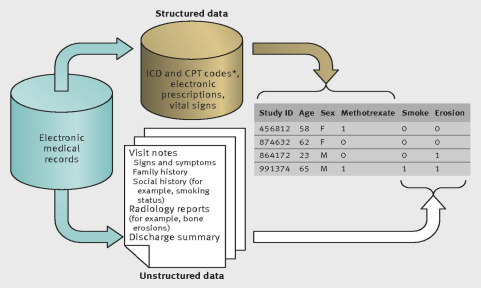
nature REVIEWS GENETICS

Review Article | Published: 02 May 2012

Mining electronic health records: towards better research applications and clinical care

Peter B. Jensen, Lars J. Jensen & Søren Brunak

Nature Reviews Genetics **13**, 395–405 (2012)



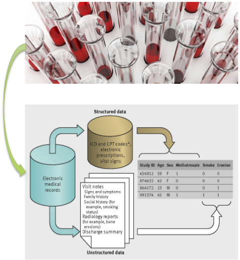
Study ID	Age	Sex	Methotrexate	Smoke	Erosion
456812	58	F	1	0	0
874632	62	F	0	0	0
864172	23	M	0	0	1
991374	65	M	1	1	1

- **Detailed** clinical and phenotypic data collected electronically for large patient cohorts, as part of routine health care delivery.
- Structured data: ICD codes, medications, lab tests, demographics etc.
- Unstructured text data (extracted from clinician notes via NLP): signs and symptoms, family history, social history, radiology reports etc.

EHR Data: The Promises and the Challenges

- Information on a **variety** of phenotypes (unlike usual cohort studies).
- Opens up unique opportunities for novel **integrative analyses**.

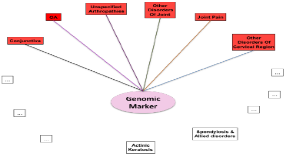
- Information on a **variety** of phenotypes (unlike usual cohort studies).
- Opens up unique opportunities for novel **integrative analyses**.



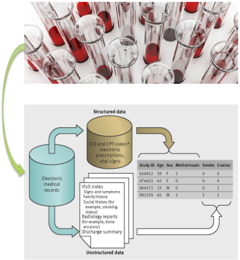
Bio-repository

EMR

- EHR + Bio-repositories $\rightsquigarrow$ genome-phenome association networks, PheWAS studies and genomic risk prediction of diseases.



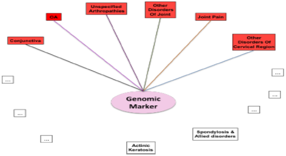
- Information on a **variety** of phenotypes (unlike usual cohort studies).
- Opens up unique opportunities for novel **integrative analyses**.



Bio-repository

EMR

- EHR + Bio-repositories $\rightsquigarrow$ genome-phenome association networks, PheWAS studies and genomic risk prediction of diseases.



- The key challenges and bottlenecks for EHR driven research:
  - Logistic difficulty in obtaining validated phenotype (Y) information.
  - Often time/labor/cost intensive (and the ICD codes are imprecise).

- Some examples of missing Y in EHRs and the reason for missingness:

EHR Data and Incompleteness: Various Examples

- Some examples of missing Y in EHRs and the reason for missingness:
  1. $Y \rightsquigarrow$ some (binary) disease phenotype (e.g. Rheumatoid Arthritis). Requires manual chart review by physicians (logistic constraints).
  2. $Y \rightsquigarrow$ some biomarker (e.g. anti-CCP, an important RA biomarker). Requires lab tests (cost constraints). Similarly, any Y requiring genomic measurements may also have cost/logistics constraints.

EHR Data and Incompleteness: Various Examples

- Some examples of missing Y in EHRs and the reason for missingness:
  1. $Y \rightsquigarrow$ some (binary) disease phenotype (e.g. Rheumatoid Arthritis). Requires manual chart review by physicians (logistic constraints).
  2. $Y \rightsquigarrow$ some biomarker (e.g. anti-CCP, an important RA biomarker). Requires lab tests (cost constraints). Similarly, any Y requiring genomic measurements may also have cost/logistics constraints.
- Verified phenotypes/treatment response/biomarkers/genomic vars (Y) available **only** for a subset. Clinical features (**X**) available for **all**.
- Further issues: selection bias/treatment by indication/preferential labeling (e.g. sicker patients get labeled/treated/tested more often).

EHR Data and Incompleteness: Various Examples

- Some examples of missing Y in EHRs and the reason for missingness:

  1. $Y \rightsquigarrow$ some (binary) disease phenotype (e.g. Rheumatoid Arthritis). Requires manual chart review by physicians (logistic constraints).
  2. $Y \rightsquigarrow$ some biomarker (e.g. anti-CCP, an important RA biomarker). Requires lab tests (cost constraints). Similarly, any Y requiring genomic measurements may also have cost/logistics constraints.

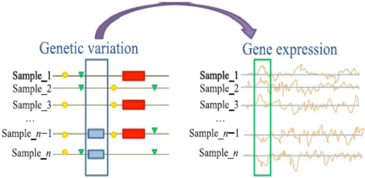
- Verified phenotypes/treatment response/biomarkers/genomic vars (Y) available **only** for a subset. Clinical features ($\mathbf{X}$) available for **all**.

- Further issues: selection bias/treatment by indication/preferential labeling (e.g. sicker patients get labeled/treated/tested more often).

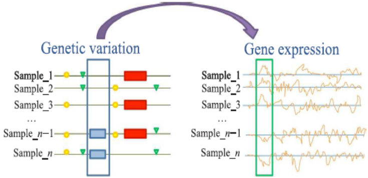
- Causal inference problems (treatment effects estimation): EHRs also facilitate comparative effectiveness research on a large scale.

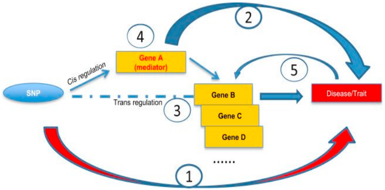
  - Many treatments/medications (and responses) being observed. All other clinical features ($\mathbf{X}$) serve as potential confounders.

- Association studies for gene expression (Y) vs. genetic variants ($\mathbf{X}$).



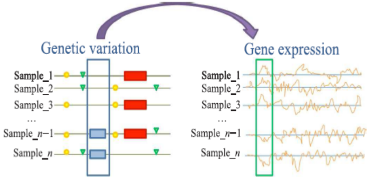
- Association studies for gene expression (Y) vs. genetic variants ($\mathbf{X}$).

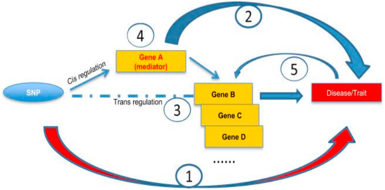




- Popular tools in integrative genomics (genetic association studies + gene expression profiling) for understanding gene regulatory networks.

- Association studies for gene expression (Y) vs. genetic variants ($\mathbf{X}$).



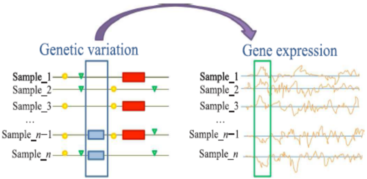


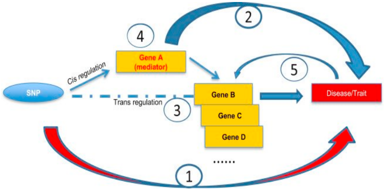
- Popular tools in integrative genomics (genetic association studies + gene expression profiling) for understanding gene regulatory networks.

- Missing data issue: gene expression data often missing (loss of power), while genetic variants data often available for a much larger group.

Another Example: eQTL Studies (Integrative Genomics)

- Association studies for gene expression (Y) vs. genetic variants ($\mathbf{X}$).





- Popular tools in integrative genomics (genetic association studies + gene expression profiling) for understanding gene regulatory networks.

- Missing data issue: gene expression data often missing (loss of power), while genetic variants data often available for a much larger group.

- Causal inference: estimate the causal effect of any one variant (the 'treatment') on Y while all other variants are potential confounders.

- **Goal for M-estimation:** estimation **and** inference, based on $\mathcal{D}_n$, of $\boldsymbol{\theta}_0 \in \mathbb{R}^d$ (possibly high dimensional), defined as the risk minimizer:

  $$\boldsymbol{\theta}_0 \equiv \boldsymbol{\theta}_0(\mathbb{P}) := \underset{\boldsymbol{\theta} \in \mathbb{R}^d}{\arg\min}\ R(\boldsymbol{\theta}), \quad \text{where } R(\boldsymbol{\theta}) := \mathbb{E}\{L(Y, \mathbf{X}, \boldsymbol{\theta})\} \text{ and}$$

  $L(\cdot) \in \mathbb{R}^+$ is any 'loss' function that is convex and differentiable in $\boldsymbol{\theta}$. Existence of $\boldsymbol{\theta}_0$ implicitly assumed (guaranteed for most usual probs).

- d can diverge with n (including $d \gg n$). Also, $\boldsymbol{\theta}_0(\mathbb{P})$ is 'model free' (no restrictions on $\mathbb{P}$). In particular, **no** model assumptions on $Y|\mathbf{X}$.

- **Goal for M-estimation:** estimation **and** inference, based on $\mathcal{D}_n$, of $\boldsymbol{\theta}_0 \in \mathbb{R}^d$ (possibly high dimensional), defined as the risk minimizer:

  $$\boldsymbol{\theta}_0 \equiv \boldsymbol{\theta}_0(\mathbb{P}) := \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^d} R(\boldsymbol{\theta}), \text{ where } R(\boldsymbol{\theta}) := \mathbb{E}\{L(Y, \mathbf{X}, \boldsymbol{\theta})\} \text{ and}$$

  $L(\cdot) \in \mathbb{R}^+$ is any 'loss' function that is convex and differentiable in $\boldsymbol{\theta}$. Existence of $\boldsymbol{\theta}_0$ implicitly assumed (guaranteed for most usual probs).

- d can diverge with n (including $d \gg n$). Also, $\boldsymbol{\theta}_0(\mathbb{P})$ is 'model free' (no restrictions on $\mathbb{P}$). In particular, **no** model assumptions on $Y|\mathbf{X}$.

- The **key challenges:** the missingness via T (if not accounted for, the estimator **will** be inconsistent!) and the high dimensional setting.

- Need suitable methods - involves estimation of nuisance functions and careful analyses (due to error terms with complex dependencies).

- **Goal for M-estimation:** estimation **and** inference, based on $\mathcal{D}_n$, of $\boldsymbol{\theta}_0 \in \mathbb{R}^d$ (possibly high dimensional), defined as the risk minimizer:

$$\boldsymbol{\theta}_0 \equiv \boldsymbol{\theta}_0(\mathbb{P}) := \underset{\boldsymbol{\theta} \in \mathbb{R}^d}{\arg\min}\, R(\boldsymbol{\theta}), \quad \text{where } R(\boldsymbol{\theta}) := \mathbb{E}\{L(Y, \mathbf{X}, \boldsymbol{\theta})\} \text{ and}$$

  $L(\cdot) \in \mathbb{R}^+$ is any 'loss' function that is convex and differentiable in $\boldsymbol{\theta}$. Existence of $\boldsymbol{\theta}_0$ implicitly assumed (guaranteed for most usual probs).

- d can diverge with n (including $d \gg n$). Also, $\boldsymbol{\theta}_0(\mathbb{P})$ is 'model free' (no restrictions on $\mathbb{P}$). In particular, **no** model assumptions on $Y | \mathbf{X}$.

- The **key challenges:** the missingness via T (if not accounted for, the estimator **will** be inconsistent!) and the high dimensional setting.

- Need suitable methods - involves estimation of nuisance functions and careful analyses (due to error terms with complex dependencies).

- Special (but low-d) case: $\boldsymbol{\theta}_0 = \mathbb{E}(Y)$ and $L(Y, \mathbf{X}, \boldsymbol{\theta}) = (Y - \boldsymbol{\theta})^2$. Leads to the average treatment effect (ATE) estimation prob in CI.

- The framework includes a broad class of M/Z-estimation problems.
- M-estimation for fully observed data: well studied with rich literature. Classical settings: Van der Vaart (2000); High dimensional settings: Negahban et al. (2012), Loh and Wainwright (2012, 2015) etc.

- The framework includes a broad class of M/Z-estimation problems.
- M-estimation for fully observed data: well studied with rich literature. Classical settings: Van der Vaart (2000); High dimensional settings: Negahban et al. (2012), Loh and Wainwright (2012, 2015) etc.
- Missing data/causal inference problems: semi-parametric inference.
  - Classical settings: vast literature (typically for mean estimation). Tsiatis (2007); Bang and Robins (2005); Robins et al. (1994) etc.

M-Estimation and Missing Data/Causal Inference Problems: A Review

- The framework includes a broad class of M/Z-estimation problems.
- M-estimation for fully observed data: well studied with rich literature. Classical settings: Van der Vaart (2000); High dimensional settings: Negahban et al. (2012), Loh and Wainwright (2012, 2015) etc.
- Missing data/causal inference problems: semi-parametric inference.
  - Classical settings: vast literature (typically for mean estimation). Tsiatis (2007); Bang and Robins (2005); Robins et al. (1994) etc.
  - High dimensional settings (but low dimensional parameters): lot of attention in recent times on mean (or ATE) estimation. Belloni et al. (2014, 2017); Farrell (2015); Chernozhukov et al. (2018).

- The framework includes a broad class of M/Z-estimation problems.
- *M*-estimation for fully observed data: well studied with rich literature. Classical settings: Van der Vaart (2000); High dimensional settings: Negahban et al. (2012), Loh and Wainwright (2012, 2015) etc.
- Missing data/causal inference problems: semi-parametric inference.
  - Classical settings: vast literature (typically for mean estimation). Tsiatis (2007); Bang and Robins (2005); Robins et al. (1994) etc.
  - High dimensional settings (but low dimensional parameters): lot of attention in recent times on mean (or ATE) estimation. Belloni et al. (2014, 2017); Farrell (2015); Chernozhukov et al. (2018).
- Much less attention when the parameter **itself** is high dimensional.

M-Estimation and Missing Data/Causal Inference Problems: A Review

- The framework includes a broad class of M/Z-estimation problems.
- M-estimation for fully observed data: well studied with rich literature. Classical settings: Van der Vaart (2000); High dimensional settings: Negahban et al. (2012), Loh and Wainwright (2012, 2015) etc.
- Missing data/causal inference problems: semi-parametric inference.
  - Classical settings: vast literature (typically for mean estimation). Tsiatis (2007); Bang and Robins (2005); Robins et al. (1994) etc.
  - High dimensional settings (but low dimensional parameters): lot of attention in recent times on mean (or ATE) estimation. Belloni et al. (2014, 2017); Farrell (2015); Chernozhukov et al. (2018).
- Much less attention when the parameter **itself** is high dimensional.
- This work contributes to **both** literature above: M-estimation + missing data + high dimensional setting **and** parameter. (Also has applications in heterogeneous treatment effects estimation in CI).

HD M-Estimation: A Few (Class of) Applications

1. All standard high dimensional (HD) regression problems with: (a) missing outcomes and (b) potentially misspecified (working) models.

1. All standard high dimensional (HD) regression problems with: (a) missing outcomes and (b) potentially misspecified (working) models.

   - E.g. squared loss: $L(Y, \mathbf{X}, \boldsymbol{\theta}) := (Y - \mathbf{X}'\boldsymbol{\theta})^2 \rightsquigarrow$ linear regression; logistic loss: $L(Y, \mathbf{X}, \boldsymbol{\theta}) := \log\{1 + \exp(\mathbf{X}'\boldsymbol{\theta})\} - Y(\mathbf{X}'\boldsymbol{\theta}) \rightsquigarrow$ logistic regression (for binary Y), exponential loss (Poisson reg.) so on . . ..
   - Note: throughout, **regardless** of any motivating 'working model' being true or not, the definition of $\boldsymbol{\theta}_0$ is completely 'model free'.

1. All standard high dimensional (HD) regression problems with: (a) missing outcomes and (b) potentially misspecified (working) models.
   - E.g. squared loss: $L(Y, \mathbf{X}, \boldsymbol{\theta}) := (Y - \mathbf{X}'\boldsymbol{\theta})^2 \rightsquigarrow$ linear regression; logistic loss: $L(Y, \mathbf{X}, \boldsymbol{\theta}) := \log\{1 + \exp(\mathbf{X}'\boldsymbol{\theta})\} - Y(\mathbf{X}'\boldsymbol{\theta}) \rightsquigarrow$ logistic regression (for binary Y), exponential loss (Poisson reg.) so on ....
   - Note: throughout, **regardless** of any motivating 'working model' being true or not, the definition of $\boldsymbol{\theta}_0$ is completely 'model free'.
2. Series estimation problems (model free) with missing Y and HD basis functions (instead of $\mathbf{X}$ in Example 1 above). E.g. spline bases.
   - Use the same choices of $L(\cdot)$ as in Example 1 above with $\mathbf{X}$ replaced by any set of d (possibly HD) basis functions $\boldsymbol{\Psi}(\mathbf{X}) := \{\psi_j(\mathbf{X})\}_{j=1}^{d}$.
   - E.g. polynomial bases: $\boldsymbol{\Psi}(\mathbf{X}) := \{1, \mathbf{x}_j^k : 1 \leq j \leq p, 1 \leq k \leq d_0\}$. ($d_0 = 1 \rightsquigarrow$ linear bases as in Example 1; $d_0 = 3 \rightsquigarrow$ cubic splines).

Another Application: HD Single Index Models (SIMs)

- Signal recovery in high dimensional single index models (SIMs) with elliptically symmetric design distribution (e.g. $\mathbf{X}$ is Gaussian).

- Let $Y = f(\beta_0'\mathbf{X}, \epsilon)$ with $f : \mathbb{R}^2 \to \mathcal{Y}$ **unknown** (i.e. $\boldsymbol{\beta}_0$ identifiable **only** upto scalar multiples) and $\epsilon \perp\!\!\!\perp \mathbf{X}$ (i.e., $Y \perp\!\!\!\perp \mathbf{X} \mid \beta_0'\mathbf{X}$).

- Signal recovery in high dimensional single index models (SIMs) with elliptically symmetric design distribution (e.g. $\mathbf{X}$ is Gaussian).
- Let $Y = f(\beta_0'\mathbf{X}, \epsilon)$ with $f : \mathbb{R}^2 \rightarrow \mathcal{Y}$ **unknown** (i.e. $\boldsymbol{\beta}_0$ identifiable **only** upto scalar multiples) and $\epsilon \perp\!\!\!\perp \mathbf{X}$ (i.e., $Y \perp\!\!\!\perp \mathbf{X} \mid \beta_0'\mathbf{X}$).
- Consider **any** of the regression problems introduced in Example 1.
  - Let $\boldsymbol{\theta}_0 := \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^p} \mathbb{E}\{L(Y, \mathbf{X}'\boldsymbol{\theta})\}$ for any convex loss function $L(\cdot) : \mathbb{R}^2 \rightarrow \mathbb{R}$ (convex in the second argument). Then, $\boldsymbol{\theta}_0 \propto \boldsymbol{\beta}_0$!
  - A remarkable result due to Li and Duan (1989).

- Signal recovery in high dimensional single index models (SIMs) with elliptically symmetric design distribution (e.g. $\mathbf{X}$ is Gaussian).
- Let $Y = f(\beta_0'\mathbf{X}, \epsilon)$ with $f : \mathbb{R}^2 \to \mathcal{Y}$ **unknown** (i.e. $\boldsymbol{\beta}_0$ identifiable **only** upto scalar multiples) and $\epsilon \perp\!\!\!\perp \mathbf{X}$ (i.e., $Y \perp\!\!\!\perp \mathbf{X} \mid \beta_0'\mathbf{X}$).
- Consider **any** of the regression problems introduced in Example 1.
  - Let $\boldsymbol{\theta}_0 := \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^p} \mathbb{E}\{L(Y, \mathbf{X}'\boldsymbol{\theta})\}$ for any convex loss function $L(\cdot) : \mathbb{R}^2 \to \mathbb{R}$ (convex in the second argument). Then, $\boldsymbol{\theta}_0 \propto \boldsymbol{\beta}_0$!
  - A remarkable result due to Li and Duan (1989).
- **Classic example** of a misspecified parametric model defining $\boldsymbol{\theta}_0$, yet $\boldsymbol{\theta}_0$ directly relates to an actual (interpretable) semi-parametric model!
  - The proportionality result also preserves any sparsity assumptions.

Applications in Causal Inference (Treatment Effects Estimation)

- Applications of all these problems in causal inference (estimation of treatment effects with useful applications in precision medicine):

- Applications of all these problems in causal inference (estimation of treatment effects with useful applications in precision medicine):

  1. Linear heterogeneous treatment effects estimation: application of the linear regression example (twice). Write $\{Y_{(0)}, Y_{(1)}\}$ linearly as:

$$Y_{(j)} = \mathbf{X}'\boldsymbol{\beta}_{(j)} + \epsilon_{(j)}, \ \ \mathbb{E}(\epsilon_{(j)}\mathbf{X}) = \mathbf{0} \ \forall\, j = 0, 1, \ \text{ so that}$$
$$Y_{(1)} - Y_{(0)} = \mathbf{X}'\boldsymbol{\beta}^* + \epsilon^*, \ \ \boldsymbol{\beta}^* := \boldsymbol{\beta}_{(1)} - \boldsymbol{\beta}_{(0)}, \ \epsilon^* := \epsilon_{(1)} - \epsilon_{(0)}.$$

- Applications of all these problems in causal inference (estimation of treatment effects with useful applications in precision medicine):

  1. Linear heterogeneous treatment effects estimation: application of the linear regression example (twice). Write $\{Y_{(0)}, Y_{(1)}\}$ linearly as:

$$Y_{(j)} = \mathbf{X}'\boldsymbol{\beta}_{(j)} + \epsilon_{(j)}, \ \ \mathbb{E}(\epsilon_{(j)}\mathbf{X}) = \mathbf{0} \ \forall\, j = 0, 1, \ \ \text{so that}$$
$$Y_{(1)} - Y_{(0)} = \mathbf{X}'\boldsymbol{\beta}^* + \epsilon^*, \ \ \boldsymbol{\beta}^* := \boldsymbol{\beta}_{(1)} - \boldsymbol{\beta}_{(0)}, \ \epsilon^* := \epsilon_{(1)} - \epsilon_{(0)}.$$

  $\boldsymbol{\beta}^*$ denotes the (model free) linear projection of $Y_{(1)} - Y_{(0)} | \mathbf{X}$. Of interest in HD settings when $\mathbb{E}\{Y_{(1)} - Y_{(0)} | \mathbf{X}\}$ is difficult to model (Chernozhukov et al., 2017; Chernozhukov and Semenova, 2017).

- Applications of all these problems in causal inference (estimation of treatment effects with useful applications in precision medicine):

  1. Linear heterogeneous treatment effects estimation: application of the linear regression example (twice). Write $\{Y_{(0)}, Y_{(1)}\}$ linearly as:

     $$Y_{(j)} = \mathbf{X}'\beta_{(j)} + \epsilon_{(j)}, \;\; \mathbb{E}(\epsilon_{(j)}\mathbf{X}) = \mathbf{0} \;\; \forall\, j = 0, 1, \;\; \text{so that}$$
     $$Y_{(1)} - Y_{(0)} = \mathbf{X}'\beta^* + \epsilon^*, \;\; \beta^* := \beta_{(1)} - \beta_{(0)}, \;\; \epsilon^* := \epsilon_{(1)} - \epsilon_{(0)}.$$

     β^* denotes the (model free) linear projection of $Y_{(1)} - Y_{(0)} | \mathbf{X}$. Of interest in HD settings when $\mathbb{E}\{Y_{(1)} - Y_{(0)} | \mathbf{X}\}$ is difficult to model (Chernozhukov et al., 2017; Chernozhukov and Semenova, 2017).

  2. Average conditional treatment effects (ACTE) estimation via series estimators: application of the series estimation example (twice).

  3. Causal inference via SIMs (signal recovery, ACTE estimation and ATE estimation): application of the SIM example (twice).

- Some notations: $m(\mathbf{X}) := \mathbb{E}(Y|\mathbf{X})$ and $\phi(\mathbf{X}, \boldsymbol{\theta}) := \mathbb{E}\{L(Y, \mathbf{X}, \boldsymbol{\theta})|\mathbf{X}\}$.

- Some notations: $m(\mathbf{X}) := \mathbb{E}(Y|\mathbf{X})$ and $\phi(\mathbf{X}, \boldsymbol{\theta}) := \mathbb{E}\{L(Y, \mathbf{X}, \boldsymbol{\theta})|\mathbf{X}\}$.
- It is generally necessary to 'account' for the missingness in Y. The 'complete case' estimator of $\boldsymbol{\theta}_0$ in general will be **inconsistent**!

- Some notations: $m(\mathbf{X}) := \mathbb{E}(Y|\mathbf{X})$ and $\phi(\mathbf{X}, \boldsymbol{\theta}) := \mathbb{E}\{L(Y, \mathbf{X}, \boldsymbol{\theta})|\mathbf{X}\}$.
- It is generally necessary to 'account' for the missingness in Y. The 'complete case' estimator of $\boldsymbol{\theta}_0$ in general will be **inconsistent**!
  - That estimator may be consistent **only if:** (1) $\nabla\phi(\mathbf{X}, \boldsymbol{\theta}_0) = \mathbf{0}$ a.s. for **every X** (for regression problems, this indicates the 'correct model' case), and/or (2) $T \perp\!\!\!\perp (Y, X)$ (i.e. the MCAR case).
  - Illustration of (1) for sq. loss: $\nabla\phi(\mathbf{X}, \boldsymbol{\theta}_0) = \mathbb{E}\{\mathbf{X}(Y - \mathbf{X}'\boldsymbol{\theta}_0)|\mathbf{X}\} = \mathbf{0}$. Hence, $\mathbb{E}(Y|\mathbf{X}) = \mathbf{X}'\boldsymbol{\theta}_0$ (i.e. a 'linear model' holds for $Y|\mathbf{X}$).

- Some notations: $m(\mathbf{X}) := \mathbb{E}(Y|\mathbf{X})$ and $\phi(\mathbf{X}, \boldsymbol{\theta}) := \mathbb{E}\{L(Y, \mathbf{X}, \boldsymbol{\theta})|\mathbf{X}\}$.
- It is generally necessary to 'account' for the missingness in Y. The 'complete case' estimator of $\boldsymbol{\theta}_0$ in general will be **inconsistent**!
  - That estimator may be consistent **only if:** (1) $\boldsymbol{\nabla}\phi(\mathbf{X}, \boldsymbol{\theta}_0) = \mathbf{0}$ a.s. for **every X** (for regression problems, this indicates the 'correct model' case), and/or (2) $T \perp\!\!\!\perp (Y, X)$ (i.e. the MCAR case).
  - Illustration of (1) for sq. loss: $\boldsymbol{\nabla}\phi(\mathbf{X}, \boldsymbol{\theta}_0) = \mathbb{E}\{\mathbf{X}(Y - \mathbf{X}'\boldsymbol{\theta}_0)|\mathbf{X}\} = \mathbf{0}$. Hence, $\mathbb{E}(Y|\mathbf{X}) = \mathbf{X}'\boldsymbol{\theta}_0$ (i.e. a 'linear model' holds for $Y|\mathbf{X}$).
- With $\boldsymbol{\theta}_0$ (and $\mathbf{X}$) being high dimensional (compared to n), we **need** some further structural constraints on $\boldsymbol{\theta}_0$ to estimate it using $\mathcal{D}_n$.
  - We **assume** that $\boldsymbol{\theta}_0$ is s-sparse: $\|\boldsymbol{\theta}_0\|_0 := s$ and $s \leq \min(n, d)$.
  - Note: the sparsity requirement has attractive (and fairly intuitive) geometric justification for **all the examples** we have given here.

- Under MAR assmpn., $R(\boldsymbol{\theta}) := \mathbb{E}\{L(Y, \mathbf{X}, \boldsymbol{\theta})\} \equiv \mathbb{E}_{\mathbf{X}}\{\phi(\mathbf{X}, \boldsymbol{\theta})\}$ admits the following debiased and doubly robust (DDR) representation:

- Under MAR assmpn., $R(\boldsymbol{\theta}) := \mathbb{E}\{L(Y, \mathbf{X}, \boldsymbol{\theta})\} \equiv \mathbb{E}_{\mathbf{X}}\{\phi(\mathbf{X}, \boldsymbol{\theta})\}$ admits the following debiased and doubly robust (DDR) representation:

$$R(\boldsymbol{\theta}) \; = \; \mathbb{E}_{\mathbf{X}}\{\phi(\mathbf{X}, \boldsymbol{\theta})\} + \mathbb{E}\left[\frac{T}{\pi(\mathbf{X})}\{L(Y, \mathbf{X}, \boldsymbol{\theta}) - \phi(\mathbf{X}, \boldsymbol{\theta})\}\right]. \quad (1)$$

  Purely non-parametric identification based on the observable $\mathbf{Z}$ and the nuisance functions: $\pi(\mathbf{X})$ and $\phi(\mathbf{X}, \boldsymbol{\theta})$ (unknown but **estimable**).

- Under MAR assmpn., $R(\boldsymbol{\theta}) := \mathbb{E}\{L(Y, \mathbf{X}, \boldsymbol{\theta})\} \equiv \mathbb{E}_{\mathbf{X}}\{\phi(\mathbf{X}, \boldsymbol{\theta})\}$ admits the following debiased and doubly robust (DDR) representation:

$$R(\boldsymbol{\theta}) \; = \; \mathbb{E}_{\mathbf{X}}\{\phi(\mathbf{X}, \boldsymbol{\theta})\} + \mathbb{E}\left[\frac{T}{\pi(\mathbf{X})}\{L(Y, \mathbf{X}, \boldsymbol{\theta}) - \phi(\mathbf{X}, \boldsymbol{\theta})\}\right]. \qquad (1)$$

  Purely non-parametric identification based on the observable $\mathbf{Z}$ and the nuisance functions: $\pi(\mathbf{X})$ and $\phi(\mathbf{X}, \boldsymbol{\theta})$ (unknown but **estimable**).

  - 2^{nd} term is simply 0, can be seen as a 'debiasing' term (of sorts).
    - Plays a **crucial** role in analyzing the empirical version of (1). Ensures first order insensitivity to any estimation errors of $\pi(\cdot)$ and $\phi(\cdot)$.

- Under MAR assmpn., $R(\boldsymbol{\theta}) := \mathbb{E}\{L(Y, \mathbf{X}, \boldsymbol{\theta})\} \equiv \mathbb{E}_{\mathbf{X}}\{\phi(\mathbf{X}, \boldsymbol{\theta})\}$ admits the following debiased and doubly robust (DDR) representation:

$$R(\boldsymbol{\theta}) \;=\; \mathbb{E}_{\mathbf{X}}\{\phi(\mathbf{X}, \boldsymbol{\theta})\} + \mathbb{E}\left[\frac{T}{\pi(\mathbf{X})}\{L(Y, \mathbf{X}, \boldsymbol{\theta}) - \phi(\mathbf{X}, \boldsymbol{\theta})\}\right]. \qquad (1)$$

  Purely non-parametric identification based on the observable $\mathbf{Z}$ and the nuisance functions: $\pi(\mathbf{X})$ and $\phi(\mathbf{X}, \boldsymbol{\theta})$ (unknown but **estimable**).

  - 2^{nd} term is simply 0, can be seen as a 'debiasing' term (of sorts).
    - Plays a **crucial** role in analyzing the empirical version of (1). Ensures first order insensitivity to any estimation errors of $\pi(\cdot)$ and $\phi(\cdot)$.
  - Double robustness (DR) aspect: replace $\{\phi(\mathbf{X}, \boldsymbol{\theta}), \pi(\mathbf{X})\}$ by **any** $\{\phi^*(\mathbf{X}, \boldsymbol{\theta}), \pi^*(\mathbf{X})\}$ and (1) **continues to hold** as long as one but **not** necessarily both of $\phi^*(\cdot) = \phi(\cdot)$ or $\pi^*(\cdot) = \pi(\cdot)$ hold.

- Given **any** estimators $\{\widehat{\pi}(\cdot), \widehat{\phi}(\cdot)\}$ be of the nuisance fns. $\{\pi(\cdot), \phi(\cdot)\}$, we define our L_1-penalized DDR estimator $\widehat{\boldsymbol{\theta}}_{\mathrm{DDR}}$ of $\boldsymbol{\theta}_0$ as:

$$\widehat{\boldsymbol{\theta}}_{\mathrm{DDR}} \equiv \widehat{\boldsymbol{\theta}}_{\mathrm{DDR}}(\lambda_n) := \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \left\{ \mathcal{L}_n^{\mathrm{DDR}}(\boldsymbol{\theta}) + \lambda_n \|\boldsymbol{\theta}\|_1 \right\}, \text{ where}$$

$$\mathcal{L}_n^{\mathrm{DDR}}(\boldsymbol{\theta}) := \frac{1}{n} \sum_{i=1}^{n} \widehat{\phi}(\mathbf{X}_i, \boldsymbol{\theta}) + \frac{T_i}{\widehat{\pi}(\mathbf{X}_i)} \left\{ L(Y_i, \mathbf{X}_i, \boldsymbol{\theta}) - \widehat{\phi}(\mathbf{X}_i, \boldsymbol{\theta}) \right\},$$

  $\lambda_n \geq 0$ is the tuning parameter and $\{\widehat{\pi}(\cdot), \widehat{\phi}(\cdot)\}$ are arbitrary except for satisfying two **basic conditions** regarding their construction:

The DDR Estimator of $\boldsymbol{\theta}_0$

- Given **any** estimators $\{\widehat{\pi}(\cdot), \widehat{\phi}(\cdot)\}$ be of the nuisance fns. $\{\pi(\cdot), \phi(\cdot)\}$, we define our L_1-penalized DDR estimator $\widehat{\boldsymbol{\theta}}_{\text{DDR}}$ of $\boldsymbol{\theta}_0$ as:

$$\widehat{\boldsymbol{\theta}}_{\text{DDR}} \equiv \widehat{\boldsymbol{\theta}}_{\text{DDR}}(\lambda_n) := \underset{\boldsymbol{\theta} \in \mathbb{R}^d}{\arg\min} \left\{ \mathcal{L}_n^{\text{DDR}}(\boldsymbol{\theta}) + \lambda_n \|\boldsymbol{\theta}\|_1 \right\}, \text{ where}$$

$$\mathcal{L}_n^{\text{DDR}}(\boldsymbol{\theta}) := \frac{1}{n} \sum_{i=1}^{n} \widehat{\phi}(\mathbf{X}_i, \boldsymbol{\theta}) + \frac{T_i}{\widehat{\pi}(\mathbf{X}_i)} \left\{ L(Y_i, \mathbf{X}_i, \boldsymbol{\theta}) - \widehat{\phi}(\mathbf{X}_i, \boldsymbol{\theta}) \right\},$$

  $\lambda_n \geq 0$ is the tuning parameter and $\{\widehat{\pi}(\cdot), \widehat{\phi}(\cdot)\}$ are arbitrary except for satisfying two **basic conditions** regarding their construction:

  - $\widehat{\pi}(\cdot)$ obtained from the data $\mathcal{T}_n := \{T_i, \mathbf{X}_i\}_{i=1}^n$ **only**; $\{\widehat{\phi}(\mathbf{X}_i, \boldsymbol{\theta})\}_{i=1}^n$ obtained in a **'cross-fitted' manner** (via sample splitting).

- Assume (temporarily) $\{\widehat{\pi}(\cdot), \widehat{\phi}(\cdot)\}$ are **both** 'correct'. DR properties (consistency) of $\widehat{\boldsymbol{\theta}}_{\text{DDR}}$ under their misspecfications discussed later.

Simplifying Assumptions and User Friendly Implementation Algorithm

- **For simplicity,** assume that the gradient $\nabla L(Y, \mathbf{X}, \boldsymbol{\theta})$ of $L(\cdot)$ satisfies a 'separable form' as follows: for some $\mathbf{h}(\mathbf{X}) \in \mathbb{R}^d$ and $g(\mathbf{X}, \boldsymbol{\theta}) \in \mathbb{R}$,

- **For simplicity,** assume that the gradient $\nabla L(Y, \mathbf{X}, \boldsymbol{\theta})$ of $L(\cdot)$ satisfies a 'separable form' as follows: for some $\mathbf{h}(\mathbf{X}) \in \mathbb{R}^d$ and $g(\mathbf{X}, \boldsymbol{\theta}) \in \mathbb{R}$,

  $$\nabla L(Y, \mathbf{X}, \boldsymbol{\theta}) = \mathbf{h}(\mathbf{X})\{Y - g(\mathbf{X}, \boldsymbol{\theta})\}, \text{ and hence,}$$
  $$\nabla \widehat{\phi}(\mathbf{X}, \boldsymbol{\theta}) = \mathbf{h}(\mathbf{X})\{\widehat{m}(\mathbf{X}) - g(\mathbf{X}, \boldsymbol{\theta})\}, \text{ where}$$

  $\widehat{m}(\mathbf{X})$ denotes the corresponding (cross-fitted) estimator of $m(\mathbf{X})$. This simplifying assumption **holds** for **all** examples given before.

  - Assumed form $\Rightarrow$ only need to obtain $\widehat{m}(\mathbf{X}_i)$ and not $\widehat{\phi}(\mathbf{X}_i, \boldsymbol{\theta})$.

- **For simplicity,** assume that the gradient $\nabla L(Y, \mathbf{X}, \boldsymbol{\theta})$ of $L(\cdot)$ satisfies a 'separable form' as follows: for some $\mathbf{h}(\mathbf{X}) \in \mathbb{R}^d$ and $g(\mathbf{X}, \boldsymbol{\theta}) \in \mathbb{R}$,

$$\nabla L(Y, \mathbf{X}, \boldsymbol{\theta}) = \mathbf{h}(\mathbf{X})\{Y - g(\mathbf{X}, \boldsymbol{\theta})\}, \quad \text{and hence,}$$
$$\nabla \widehat{\phi}(\mathbf{X}, \boldsymbol{\theta}) = \mathbf{h}(\mathbf{X})\{\widehat{m}(\mathbf{X}) - g(\mathbf{X}, \boldsymbol{\theta})\}, \quad \text{where}$$

  $\widehat{m}(\mathbf{X})$ denotes the corresponding (cross-fitted) estimator of $m(\mathbf{X})$. This simplifying assumption **holds** for **all** examples given before.

  - Assumed form $\Rightarrow$ only need to obtain $\widehat{m}(\mathbf{X}_i)$ and not $\widehat{\phi}(\mathbf{X}_i, \boldsymbol{\theta})$.

- **Implementation algorithm.** $\widehat{\boldsymbol{\theta}}_{\text{DDR}}$ can be obtained **simply** as:

$$\widehat{\boldsymbol{\theta}}_{\text{DDR}} \equiv \widehat{\boldsymbol{\theta}}_{\text{DDR}}(\lambda_n) := \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \left\{ \frac{1}{n} \sum_{i=1}^{n} L(\widetilde{Y}_i, \mathbf{X}_i, \boldsymbol{\theta}) + \lambda_n \|\boldsymbol{\theta}\|_1 \right\},$$

  where $\widetilde{Y}_i := \widehat{m}(\mathbf{X}_i) + \frac{T_i}{\widehat{\pi}(\mathbf{X}_i)}\{Y_i - \widehat{m}(\mathbf{X}_i)\}$, $\forall\, i$, is a 'pseudo' outcome. Can use '`glmnet`' in `R`. Pretend to have a 'full' data: $\{\widetilde{Y}_i, \mathbf{X}_i\}_{i=1}^{n}$.

- Assume $L(\cdot)$ is convex and differentiable in $\boldsymbol{\theta}$ and $\mathcal{L}_n^{\mathrm{DDR}}(\boldsymbol{\theta})$ satisfies the Restricted Strong Convexity (RSC) condition (Negahban et al., 2012) at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. Then, for any choice of $\lambda_n \geq 2 \left\| \boldsymbol{\nabla} \mathcal{L}_n^{\mathrm{DDR}}(\boldsymbol{\theta}_0) \right\|_\infty$,

- Assume $L(\cdot)$ is convex and differentiable in $\boldsymbol{\theta}$ and $\mathcal{L}_n^{\mathrm{DDR}}(\boldsymbol{\theta})$ satisfies the Restricted Strong Convexity (RSC) condition (Negahban et al., 2012) at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. Then, for any choice of $\lambda_n \geq 2 \left\| \nabla \mathcal{L}_n^{\mathrm{DDR}}(\boldsymbol{\theta}_0) \right\|_\infty$,

$$\left\| \widehat{\boldsymbol{\theta}}_{\mathrm{DDR}}(\lambda_n) - \boldsymbol{\theta}_0 \right\|_2 \lesssim \lambda_n \sqrt{s}, \text{ and } \left\| \widehat{\boldsymbol{\theta}}_{\mathrm{DDR}}(\lambda_n) - \boldsymbol{\theta}_0 \right\|_1 \lesssim \lambda_n s.$$

  where $s := \|\boldsymbol{\theta}_0\|_0$. This is a **deterministic** deviation bound. Holds for any choices of $\{\widehat{\pi}(\cdot), \widehat{m}(\cdot)\}$ and for any realization of $\mathcal{D}_n$.

- Assume $L(\cdot)$ is convex and differentiable in $\boldsymbol{\theta}$ and $\mathcal{L}_n^{\mathrm{DDR}}(\boldsymbol{\theta})$ satisfies the Restricted Strong Convexity (RSC) condition (Negahban et al., 2012) at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. Then, for any choice of $\lambda_n \geq 2\,\|\boldsymbol{\nabla}\mathcal{L}_n^{\mathrm{DDR}}(\boldsymbol{\theta}_0)\|_\infty$,

$$\left\|\widehat{\boldsymbol{\theta}}_{\mathrm{DDR}}(\lambda_n) - \boldsymbol{\theta}_0\right\|_2 \lesssim \lambda_n\sqrt{s}, \text{ and } \left\|\widehat{\boldsymbol{\theta}}_{\mathrm{DDR}}(\lambda_n) - \boldsymbol{\theta}_0\right\|_1 \lesssim \lambda_n s.$$

  where $s := \|\boldsymbol{\theta}_0\|_0$. This is a **deterministic** deviation bound. Holds for any choices of $\{\widehat{\pi}(\cdot), \widehat{m}(\cdot)\}$ and for any realization of $\mathcal{D}_n$.

  - The RSC (or 'cone') condition for $\mathcal{L}_n^{\mathrm{DDR}}(\boldsymbol{\theta})$ is exactly the **same** as the usual RSC condition required under a fully observed data! The fully observed data RSC condition's validity is well studied.

- Assume $L(\cdot)$ is convex and differentiable in $\boldsymbol{\theta}$ and $\mathcal{L}_n^{\mathrm{DDR}}(\boldsymbol{\theta})$ satisfies the Restricted Strong Convexity (RSC) condition (Negahban et al., 2012) at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. Then, for any choice of $\lambda_n \geq 2\left\|\nabla \mathcal{L}_n^{\mathrm{DDR}}(\boldsymbol{\theta}_0)\right\|_\infty$,

$$\left\|\widehat{\boldsymbol{\theta}}_{\mathrm{DDR}}(\lambda_n) - \boldsymbol{\theta}_0\right\|_2 \lesssim \lambda_n \sqrt{s}, \text{ and } \left\|\widehat{\boldsymbol{\theta}}_{\mathrm{DDR}}(\lambda_n) - \boldsymbol{\theta}_0\right\|_1 \lesssim \lambda_n s.$$

  where $s := \|\boldsymbol{\theta}_0\|_0$. This is a **deterministic** deviation bound. Holds for any choices of $\{\widehat{\pi}(\cdot), \widehat{m}(\cdot)\}$ and for any realization of $\mathcal{D}_n$.

  - The RSC (or 'cone') condition for $\mathcal{L}_n^{\mathrm{DDR}}(\boldsymbol{\theta})$ is exactly the **same** as the usual RSC condition required under a fully observed data! The fully observed data RSC condition's validity is well studied.

- Key quantity of interest: the random lower bound $\left\|\nabla \mathcal{L}_n^{\mathrm{DDR}}(\boldsymbol{\theta}_0)\right\|_\infty$ for λ_n. Need probabilistic bounds to determine convergence rate of $\widehat{\boldsymbol{\theta}}_{\mathrm{DDR}}$.

- Bounds on $\|\nabla \mathcal{L}_n^{\text{DDR}}(\boldsymbol{\theta}_0)\|_\infty$ determines the rate of choice of λ_n and hence the convergence rate of $\widehat{\boldsymbol{\theta}}_{\text{DDR}}$ (using the deviation bound).
- **Probabilistic** bounds for $\|\nabla \mathcal{L}_n^{\text{DDR}}(\boldsymbol{\theta}_0)\|_\infty$: the basic decomposition

$$\left\|\nabla \mathcal{L}_n^{\text{DDR}}(\boldsymbol{\theta}_0)\right\|_\infty \leq \|\mathbf{T}_{0,n}\|_\infty + \|\mathbf{T}_{\pi,n}\|_\infty + \|\mathbf{T}_{m,n}\|_\infty + \|\mathbf{R}_{\pi,m,n}\|_\infty ,$$

- Bounds on $\|\nabla \mathcal{L}_n^{\text{DDR}}(\boldsymbol{\theta}_0)\|_\infty$ determines the rate of choice of λ_n and hence the convergence rate of $\widehat{\boldsymbol{\theta}}_{\text{DDR}}$ (using the deviation bound).
- **Probabilistic** bounds for $\|\nabla \mathcal{L}_n^{\text{DDR}}(\boldsymbol{\theta}_0)\|_\infty$: the basic decomposition

$$\left\|\nabla \mathcal{L}_n^{\text{DDR}}(\boldsymbol{\theta}_0)\right\|_\infty \leq \|\mathbf{T}_{0,n}\|_\infty + \|\mathbf{T}_{\pi,n}\|_\infty + \|\mathbf{T}_{m,n}\|_\infty + \|\mathbf{R}_{\pi,m,n}\|_\infty ,$$

  where $\mathbf{T}_{0,n}$ is the 'main' term (a centered iid average), $\mathbf{T}_{\pi,n}$ is the 'π-error' term involving $\widehat{\pi}(\cdot) - \pi(\cdot)$ and $\mathbf{T}_{m,n}$ is the 'm-error' term involving $\widehat{m}(\cdot) - m(\cdot)$, while $\mathbf{R}_{\pi,m,n}$ is the '(π, m)-error' term (usually lower order) involving the product of $\widehat{\pi}(\cdot) - \pi(\cdot)$ and $\widehat{m}(\cdot) - m(\cdot)$.
- Control each term separately. The analyses are all non-asymptotic and nuanced, especially in order to get sharp rates for $\mathbf{T}_{\pi,n}$ and $\mathbf{T}_{m,n}$.

- Bounds on $\|\nabla \mathcal{L}_n^{\mathrm{DDR}}(\boldsymbol{\theta}_0)\|_\infty$ determines the rate of choice of λ_n and hence the convergence rate of $\widehat{\boldsymbol{\theta}}_{\mathrm{DDR}}$ (using the deviation bound).

- **Probabilistic** bounds for $\|\nabla \mathcal{L}_n^{\mathrm{DDR}}(\boldsymbol{\theta}_0)\|_\infty$: the basic decomposition

$$\left\|\nabla \mathcal{L}_n^{\mathrm{DDR}}(\boldsymbol{\theta}_0)\right\|_\infty \leq \|\mathbf{T}_{0,n}\|_\infty + \|\mathbf{T}_{\pi,n}\|_\infty + \|\mathbf{T}_{m,n}\|_\infty + \|\mathbf{R}_{\pi,m,n}\|_\infty,$$

where $\mathbf{T}_{0,n}$ is the 'main' term (a centered iid average), $\mathbf{T}_{\pi,n}$ is the 'π-error' term involving $\widehat{\pi}(\cdot) - \pi(\cdot)$ and $\mathbf{T}_{m,n}$ is the 'm-error' term involving $\widehat{m}(\cdot) - m(\cdot)$, while $\mathbf{R}_{\pi,m,n}$ is the '(π, m)-error' term (usually lower order) involving the product of $\widehat{\pi}(\cdot) - \pi(\cdot)$ and $\widehat{m}(\cdot) - m(\cdot)$.

- Control each term separately. The analyses are all non-asymptotic and nuanced, especially in order to get sharp rates for $\mathbf{T}_{\pi,n}$ and $\mathbf{T}_{m,n}$.

- We show: $\|\nabla \mathcal{L}_n^{\mathrm{DDR}}(\boldsymbol{\theta}_0)\|_\infty \lesssim \sqrt{(\log d)/n}$ with high probability, and hence $\|\widehat{\boldsymbol{\theta}}_{\mathrm{DDR}} - \boldsymbol{\theta}_0\|_2 \lesssim \sqrt{s(\log d)/n}$. So, clearly it is rate optimal.

- Basic (high level) consistency conditions on $\{\widehat{\pi}(\cdot), \widehat{m}(\cdot)\}$. Let $\{\widehat{\pi}(\cdot), \widehat{m}(\cdot)\}$ be **any** general and 'correct' estimators of $\{\pi(\cdot), m(\cdot)\}$, and assume they satisfy the following **pointwise** convergence rates:

- Basic (high level) consistency conditions on $\{\widehat{\pi}(\cdot), \widehat{m}(\cdot)\}$. Let $\{\widehat{\pi}(\cdot), \widehat{m}(\cdot)\}$ be **any** general and 'correct' estimators of $\{\pi(\cdot), m(\cdot)\}$, and assume they satisfy the following **pointwise** convergence rates:

$$|\widehat{\pi}(\mathbf{x}) - \pi(\mathbf{x})| \lesssim_{\mathbb{P}} \delta_{n,\pi} \text{ and } |\widehat{m}(\mathbf{x}) - m(\mathbf{x})| \lesssim_{\mathbb{P}} \xi_{n,m} \; \forall \, \mathbf{x} \in \mathcal{X}, \quad (2)$$

for **some** sequences $\delta_{n,\pi}, \xi_{n,m} \geq 0$ such that $(\delta_{n,\pi} + \xi_{n,m})\sqrt{\log(nd)} = o(1)$ and the product $\delta_{n,\pi}\xi_{n,m}(\log n) = o(\sqrt{(\log d)/n})$.

- Basic (high level) consistency conditions on $\{\widehat{\pi}(\cdot), \widehat{m}(\cdot)\}$. Let $\{\widehat{\pi}(\cdot), \widehat{m}(\cdot)\}$ be **any** general and 'correct' estimators of $\{\pi(\cdot), m(\cdot)\}$, and assume they satisfy the following **pointwise** convergence rates:

$$|\widehat{\pi}(\mathbf{x}) - \pi(\mathbf{x})| \lesssim_{\mathbb{P}} \delta_{n,\pi} \text{ and } |\widehat{m}(\mathbf{x}) - m(\mathbf{x})| \lesssim_{\mathbb{P}} \xi_{n,m} \; \forall \, \mathbf{x} \in \mathcal{X}, \quad (2)$$

for **some** sequences $\delta_{n,\pi}, \xi_{n,m} \geq 0$ such that $(\delta_{n,\pi} + \xi_{n,m})\sqrt{\log(nd)} = o(1)$ and the product $\delta_{n,\pi}\xi_{n,m}(\log n) = o(\sqrt{(\log d)/n})$.

- Under condition (2), along with some more 'suitable' tail assumptions (sub-Gaussian tails etc.), we have: with high probability,

$$\|\mathbf{T}_{0,n}\|_\infty \lesssim \sqrt{\frac{\log d}{n}}, \quad \|\mathbf{T}_{\pi,n}\|_\infty \lesssim \sqrt{\frac{\log d}{n}} \left\{\delta_{n,\pi}\sqrt{\log(nd)}\right\}, \quad \text{and}$$

- Basic (high level) consistency conditions on $\{\widehat{\pi}(\cdot), \widehat{m}(\cdot)\}$. Let $\{\widehat{\pi}(\cdot), \widehat{m}(\cdot)\}$ be **any** general and 'correct' estimators of $\{\pi(\cdot), m(\cdot)\}$, and assume they satisfy the following **pointwise** convergence rates:

$$|\widehat{\pi}(\mathbf{x}) - \pi(\mathbf{x})| \lesssim_{\mathbb{P}} \delta_{n,\pi} \text{ and } |\widehat{m}(\mathbf{x}) - m(\mathbf{x})| \lesssim_{\mathbb{P}} \xi_{n,m} \; \forall \, \mathbf{x} \in \mathcal{X}, \quad (2)$$

for **some** sequences $\delta_{n,\pi}, \xi_{n,m} \geq 0$ such that $(\delta_{n,\pi} + \xi_{n,m})\sqrt{\log(nd)} = o(1)$ and the product $\delta_{n,\pi}\xi_{n,m}(\log n) = o(\sqrt{(\log d)/n})$.

- Under condition (2), along with some more 'suitable' tail assumptions (sub-Gaussian tails etc.), we have: with high probability,

$$\|\mathbf{T}_{0,n}\|_\infty \lesssim \sqrt{\frac{\log d}{n}}, \quad \|\mathbf{T}_{\pi,n}\|_\infty \lesssim \sqrt{\frac{\log d}{n}} \left\{\delta_{n,\pi}\sqrt{\log(nd)}\right\}, \quad \text{and}$$

$$\|\mathbf{T}_{m,n}\|_\infty \lesssim \sqrt{\frac{\log d}{n}} \left\{\xi_{n,m}\sqrt{\log(nd)}\right\}, \quad \|\mathbf{R}_{\pi,m,n}\|_\infty \lesssim \delta_{n,\pi}\xi_{n,m}(\log n).$$

- Basic (high level) consistency conditions on $\{\widehat{\pi}(\cdot), \widehat{m}(\cdot)\}$. Let $\{\widehat{\pi}(\cdot), \widehat{m}(\cdot)\}$ be **any** general and 'correct' estimators of $\{\pi(\cdot), m(\cdot)\}$, and assume they satisfy the following **pointwise** convergence rates:

$$|\widehat{\pi}(\mathbf{x}) - \pi(\mathbf{x})| \lesssim_{\mathbb{P}} \delta_{n,\pi} \text{ and } |\widehat{m}(\mathbf{x}) - m(\mathbf{x})| \lesssim_{\mathbb{P}} \xi_{n,m} \, \forall \, \mathbf{x} \in \mathcal{X}, \quad (2)$$

for **some** sequences $\delta_{n,\pi}, \xi_{n,m} \geq 0$ such that $(\delta_{n,\pi} + \xi_{n,m})\sqrt{\log(nd)} = o(1)$ and the product $\delta_{n,\pi}\xi_{n,m}(\log n) = o(\sqrt{(\log d)/n})$.

- Under condition (2), along with some more 'suitable' tail assumptions (sub-Gaussian tails etc.), we have: with high probability,

$$\|\mathbf{T}_{0,n}\|_\infty \lesssim \sqrt{\frac{\log d}{n}}, \quad \|\mathbf{T}_{\pi,n}\|_\infty \lesssim \sqrt{\frac{\log d}{n}} \left\{\delta_{n,\pi}\sqrt{\log(nd)}\right\}, \quad \text{and}$$

$$\|\mathbf{T}_{m,n}\|_\infty \lesssim \sqrt{\frac{\log d}{n}} \left\{\xi_{n,m}\sqrt{\log(nd)}\right\}, \quad \|\mathbf{R}_{\pi,m,n}\|_\infty \lesssim \delta_{n,\pi}\xi_{n,m}(\log n).$$

- Hence, $\|\nabla\mathcal{L}_n^{\mathrm{DDR}}(\boldsymbol{\theta}_0)\|_\infty \lesssim \sqrt{\frac{\log d}{n}}\{1 + o(1)\}$ with high probability.

- Consider $\widehat{\boldsymbol{\theta}}_{\text{DDR}}$ for the squared loss: $L(Y, \mathbf{X}, \boldsymbol{\theta}) := \{Y - \boldsymbol{\Psi}(\mathbf{X})'\boldsymbol{\theta}\}^2$, where $\boldsymbol{\Psi}(\mathbf{X}) \in \mathbb{R}^d$ denotes any HD vector of basis functions of $\mathbf{X}$.

- Define $\boldsymbol{\Sigma} := \mathbb{E}\{\boldsymbol{\Psi}(\mathbf{X})\boldsymbol{\Psi}(\mathbf{X})'\}$, $\boldsymbol{\Omega} := \boldsymbol{\Sigma}^{-1}$, and let $\widehat{\boldsymbol{\Omega}}$ be any reasonable estimator of $\boldsymbol{\Omega}$ (and assume $\boldsymbol{\Omega}$ is sparse if required).

- We then define the **desparsified** DDR estimator $\widetilde{\boldsymbol{\theta}}_{\text{DDR}}$ as follows.

- Consider $\widehat{\boldsymbol{\theta}}_{\text{DDR}}$ for the squared loss: $L(Y, \mathbf{X}, \boldsymbol{\theta}) := \{Y - \boldsymbol{\Psi}(\mathbf{X})'\boldsymbol{\theta}\}^2$, where $\boldsymbol{\Psi}(\mathbf{X}) \in \mathbb{R}^d$ denotes any HD vector of basis functions of $\mathbf{X}$.

- Define $\boldsymbol{\Sigma} := \mathbb{E}\{\boldsymbol{\Psi}(\mathbf{X})\boldsymbol{\Psi}(\mathbf{X})'\}$, $\boldsymbol{\Omega} := \boldsymbol{\Sigma}^{-1}$, and let $\widehat{\boldsymbol{\Omega}}$ be any reasonable estimator of $\boldsymbol{\Omega}$ (and assume $\boldsymbol{\Omega}$ is sparse if required).

- We then define the **desparsified** DDR estimator $\widetilde{\boldsymbol{\theta}}_{\text{DDR}}$ as follows.

$$\widetilde{\boldsymbol{\theta}}_{\text{DDR}} := \widehat{\boldsymbol{\theta}}_{\text{DDR}} + \underbrace{\widehat{\boldsymbol{\Omega}} \frac{1}{n} \sum_{i=1}^{n} \{\widetilde{Y}_i - \boldsymbol{\Psi}(\mathbf{X}_i)'\widehat{\boldsymbol{\theta}}_{\text{DDR}}\}\boldsymbol{\Psi}(\mathbf{X}_i)}_{\text{Desparsification/Debiasing term}}, \quad \text{where}$$

$$\widetilde{Y}_i := \widehat{m}(\mathbf{X}_i) + \frac{T_i}{\widehat{\pi}(\mathbf{X}_i)}\{Y_i - \widehat{m}(\mathbf{X}_i)\} \quad \text{are the pseudo outcomes.}$$

- Consider $\widehat{\boldsymbol{\theta}}_{\text{DDR}}$ for the squared loss: $L(Y, \mathbf{X}, \boldsymbol{\theta}) := \{Y - \boldsymbol{\Psi}(\mathbf{X})'\boldsymbol{\theta}\}^2$, where $\boldsymbol{\Psi}(\mathbf{X}) \in \mathbb{R}^d$ denotes any HD vector of basis functions of $\mathbf{X}$.
- Define $\boldsymbol{\Sigma} := \mathbb{E}\{\boldsymbol{\Psi}(\mathbf{X})\boldsymbol{\Psi}(\mathbf{X})'\}$, $\boldsymbol{\Omega} := \boldsymbol{\Sigma}^{-1}$, and let $\widehat{\boldsymbol{\Omega}}$ be any reasonable estimator of $\boldsymbol{\Omega}$ (and assume $\boldsymbol{\Omega}$ is sparse if required).
- We then define the **desparsified** DDR estimator $\widetilde{\boldsymbol{\theta}}_{\text{DDR}}$ as follows.

$$\widetilde{\boldsymbol{\theta}}_{\text{DDR}} := \widehat{\boldsymbol{\theta}}_{\text{DDR}} + \underbrace{\widehat{\boldsymbol{\Omega}}\,\frac{1}{n}\sum_{i=1}^{n}\{\widetilde{Y}_i - \boldsymbol{\Psi}(\mathbf{X}_i)'\widehat{\boldsymbol{\theta}}_{\text{DDR}}\}\boldsymbol{\Psi}(\mathbf{X}_i)}_{\text{Desparsification/Debiasing term}}, \text{ where}$$

$$\widetilde{Y}_i := \widehat{m}(\mathbf{X}_i) + \frac{T_i}{\widehat{\pi}(\mathbf{X}_i)}\{Y_i - \widehat{m}(\mathbf{X}_i)\} \text{ are the pseudo outcomes.}$$

Debiasing similar (in spirit) to van de Geer et al. (2014), except its the 'right' one for this problem (using pseudo outcomes in the full data).

- Assume: the basic convergence conditions (2) for $\{\widehat{\pi}(\cdot), \widehat{m}(\cdot)\}$, $\boldsymbol{\Omega}\mathbf{X}$ is sub-Gaussian and that $\|\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}\|_1 = O_{\mathbb{P}}(a_n)$, $\|I - \widehat{\boldsymbol{\Omega}}\widehat{\boldsymbol{\Sigma}}\|_{\max} = O_{\mathbb{P}}(b_n)$, with $a_n\sqrt{\log d} = o(1)$ and $b_n s\sqrt{\log d} = o(1)$, where $s := \|\boldsymbol{\theta}_0\|_0$.

- Then, $\widetilde{\boldsymbol{\theta}}_{\text{DDR}}$ satisfies the asymptotic linear expansion **(ALE)**:

The Desparisified DDR Estimator: Asymptotic Linear Expansion

- Assume: the basic convergence conditions (2) for $\{\widehat{\pi}(\cdot), \widehat{m}(\cdot)\}$, $\mathbf{\Omega X}$ is sub-Gaussian and that $\|\widehat{\mathbf{\Omega}} - \mathbf{\Omega}\|_1 = O_{\mathbb{P}}(a_n)$, $\|I - \widehat{\mathbf{\Omega}}\widehat{\mathbf{\Sigma}}\|_{\max} = O_{\mathbb{P}}(b_n)$, with $a_n\sqrt{\log d} = o(1)$ and $b_n s\sqrt{\log d} = o(1)$, where $s := \|\boldsymbol{\theta}_0\|_0$.

- Then, $\widetilde{\boldsymbol{\theta}}_{\mathrm{DDR}}$ satisfies the asymptotic linear expansion **(ALE)**:

$$(\widetilde{\boldsymbol{\theta}}_{\mathrm{DDR}} - \boldsymbol{\theta}_0) \ = \ \frac{1}{n}\sum_{i=1}^{n} \mathbf{\Omega}\{\boldsymbol{\psi}_0(\mathbf{Z}_i)\} + \mathbf{\Delta}_n, \ \text{where } \|\mathbf{\Delta}_n\|_\infty = o_{\mathbb{P}}(n^{-\frac{1}{2}})$$

$$\text{and } \ \boldsymbol{\psi}_0(\mathbf{Z}) := \left[\{m(\mathbf{X}) - \mathbf{\Psi}(\mathbf{X})'\theta_0\} + \frac{T}{\pi(\mathbf{X})}\{Y - m(\mathbf{X})\}\right]\mathbf{\Psi}(\mathbf{X})$$

with $\mathbb{E}\{\boldsymbol{\psi}_0(\mathbf{Z})\} = \mathbf{0}$. The ALE facilitates inference (e.g. confidence intervals etc.) for any low-d component of $\boldsymbol{\theta}_0$ via Gaussian approx.

- Assume: the basic convergence conditions (2) for $\{\widehat{\pi}(\cdot), \widehat{m}(\cdot)\}$, $\mathbf{\Omega X}$ is sub-Gaussian and that $\|\widehat{\mathbf{\Omega}} - \mathbf{\Omega}\|_1 = O_{\mathbb{P}}(a_n)$, $\|I - \widehat{\mathbf{\Omega}}\widehat{\mathbf{\Sigma}}\|_{\max} = O_{\mathbb{P}}(b_n)$, with $a_n\sqrt{\log d} = o(1)$ and $b_n s\sqrt{\log d} = o(1)$, where $s := \|\boldsymbol{\theta}_0\|_0$.

- Then, $\widetilde{\boldsymbol{\theta}}_{\mathrm{DDR}}$ satisfies the asymptotic linear expansion **(ALE)**:

$$(\widetilde{\boldsymbol{\theta}}_{\mathrm{DDR}} - \boldsymbol{\theta}_0) \;=\; \frac{1}{n}\sum_{i=1}^{n} \mathbf{\Omega}\{\psi_0(\mathbf{Z}_i)\} + \mathbf{\Delta}_n, \text{ where } \|\mathbf{\Delta}_n\|_\infty = o_{\mathbb{P}}(n^{-\frac{1}{2}})$$

$$\text{and } \psi_0(\mathbf{Z}) := \left[\{m(\mathbf{X}) - \mathbf{\Psi}(\mathbf{X})'\theta_0\} + \frac{T}{\pi(\mathbf{X})}\{Y - m(\mathbf{X})\}\right]\mathbf{\Psi}(\mathbf{X})$$

  with $\mathbb{E}\{\psi_0(\mathbf{Z})\} = \mathbf{0}$. The ALE facilitates inference (e.g. confidence intervals etc.) for any low-d component of $\boldsymbol{\theta}_0$ via Gaussian approx.

- Further, the ALE is also 'optimal'. The function $\mathbf{\Omega}\psi_0(\mathbf{Z}) =: \mathbf{\Psi}_{\mathrm{eff}}(\mathbf{Z})$ is the 'efficient' influence function for $\boldsymbol{\theta}_0$ (Robins et al., 1994). Thus, in classical settings, $\widetilde{\boldsymbol{\theta}}_{\mathrm{DDR}}$ achieves the semi-parametric efficiency bound.

- Coordinate-wise asymptotic normality of $\widetilde{\boldsymbol{\theta}}_{\mathrm{DDR}}$: $\forall 1 \leq j \leq d$,

$$\sqrt{n}(\widetilde{\boldsymbol{\theta}}_{\mathrm{DDR}} - \boldsymbol{\theta}_0)_j \xrightarrow{d} \mathcal{N}(\mathbf{0}, \sigma_{0,j}^2), \text{ where } \sigma_{0,j}^2 := \mathrm{Var}\{\boldsymbol{\Omega}_{j\cdot}' \psi_0(\mathbf{Z})\}.$$

  Further, $\max_{1 \leq j \leq d} |\widehat{\sigma}_{0,j} - \sigma_{0,j}| = o_{\mathbb{P}}(1)$, where $\widehat{\sigma}_{0,j}$ is the plug-in estimator obtained by plugging in $\widehat{\boldsymbol{\Omega}}$, $\widehat{\pi}(\cdot)$ and $\widehat{m}(\cdot)$ in $\mathrm{Var}\{\boldsymbol{\Omega}_{j\cdot}' \psi_0(\mathbf{Z})\}$.

- Can choose $\widehat{\boldsymbol{\Omega}}$ to be **any** standard (sparse) precision matrix estimator, e.g. the node-wise Lasso estimator. Here, $a_n = s_{\Omega}\sqrt{(\log d)/n}$ and $b_n = \sqrt{(\log d)/n}$ under suitable conditions, with $s_{\boldsymbol{\Omega}} := \max_{1 \leq j \leq d} \|\boldsymbol{\Omega}_{j\cdot}\|_0$.

- Coordinate-wise asymptotic normality of $\widetilde{\boldsymbol{\theta}}_{\mathrm{DDR}}$: $\forall 1 \leq j \leq d$,

$$\sqrt{n}(\widetilde{\boldsymbol{\theta}}_{\mathrm{DDR}} - \boldsymbol{\theta}_0)_j \xrightarrow{d} \mathcal{N}(\mathbf{0}, \sigma_{0,j}^2), \text{ where } \sigma_{0,j}^2 := \mathrm{Var}\{\boldsymbol{\Omega}_{j\cdot}' \psi_0(\mathbf{Z})\}.$$

  Further, $\max_{1 \leq j \leq d} |\widehat{\sigma}_{0,j} - \sigma_{0,j}| = o_{\mathbb{P}}(1)$, where $\widehat{\sigma}_{0,j}$ is the plug-in estimator obtained by plugging in $\widehat{\boldsymbol{\Omega}}$, $\widehat{\pi}(\cdot)$ and $\widehat{m}(\cdot)$ in $\mathrm{Var}\{\boldsymbol{\Omega}_{j\cdot}' \psi_0(\mathbf{Z})\}$.

- Can choose $\widehat{\boldsymbol{\Omega}}$ to be **any** standard (sparse) precision matrix estimator, e.g. the node-wise Lasso estimator. Here, $a_n = s_\Omega \sqrt{(\log d)/n}$ and $b_n = \sqrt{(\log d)/n}$ under suitable conditions, with $s_{\boldsymbol{\Omega}} := \max\limits_{1 \leq j \leq d} \|\boldsymbol{\Omega}_{j\cdot}\|_0$.

- The error $\boldsymbol{\Delta}_n$ can be decomposed as: $\boldsymbol{\Delta}_n = \boldsymbol{\Delta}_{n,1} + \boldsymbol{\Delta}_{n,2} + \boldsymbol{\Delta}_{n,3}$, where $\boldsymbol{\Delta}_{n,1} := \frac{1}{n}(\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}) \sum_{i=1}^n \psi_0(\mathbf{Z}_i)$, $\boldsymbol{\Delta}_{n,2} := (I_d - \widehat{\boldsymbol{\Omega}}\widehat{\boldsymbol{\Sigma}})(\widehat{\boldsymbol{\theta}}_{\mathrm{DDR}} - \boldsymbol{\theta}_0)$ and $\boldsymbol{\Delta}_{n,3} := \widehat{\boldsymbol{\Omega}}(\mathbf{T}_{\pi,n} + \mathbf{T}_{m,n} + \mathbf{R}_{\pi,m,n})$, with $\|\boldsymbol{\Delta}_{n,3}\|_\infty \lesssim_{\mathbb{P}} n^{-\frac{1}{2}}$ and

$$\|\boldsymbol{\Delta}_{n,1}\|_\infty \lesssim a_n \sqrt{\frac{\log d}{n}} \text{ and } \|\boldsymbol{\Delta}_{n,2}\|_\infty \lesssim b_n s \sqrt{\frac{\log d}{n}}.$$

- Finally, let $\{\widehat{\pi}(\cdot), \widehat{m}(\cdot)\} \rightarrow \{\pi^*(\cdot), m^*(\cdot)\}$, with either $\pi^*(\cdot) = \pi(\cdot)$ or $m^*(\cdot) = m(\cdot)$ **but not** necessarily both. Assume the same pointwise convergence conditions and rates $(\delta_{n,\pi}, \xi_{n,m})$ for $\{\widehat{\pi}(\cdot), \widehat{m}(\cdot)\}$ as in (2), but now with $\{\pi(\cdot), m(\cdot)\}$ therein replaced by $\{\pi^*(\cdot), m^*(\cdot)\}$.

- Under some 'suitable' assumptions, we have: with high probability,

- Finally, let $\{\widehat{\pi}(\cdot), \widehat{m}(\cdot)\} \rightarrow \{\pi^*(\cdot), m^*(\cdot)\}$, with either $\pi^*(\cdot) = \pi(\cdot)$ or $m^*(\cdot) = m(\cdot)$ **but not** necessarily both. Assume the same pointwise convergence conditions and rates $(\delta_{n,\pi}, \xi_{n,m})$ for $\{\widehat{\pi}(\cdot), \widehat{m}(\cdot)\}$ as in (2), but now with $\{\pi(\cdot), m(\cdot)\}$ therein replaced by $\{\pi^*(\cdot), m^*(\cdot)\}$.

- Under some 'suitable' assumptions, we have: with high probability,

$$\|\mathbf{T}_{0,n}\|_\infty + \|\mathbf{T}_{\pi,n}\|_\infty + \|\mathbf{T}_{m,n}\|_\infty \lesssim \sqrt{\frac{\log d}{n}} \left\{1 + 1_{(\pi^*, m^*) \neq (\pi, m)}\right\}$$

- Finally, let $\{\widehat{\pi}(\cdot), \widehat{m}(\cdot)\} \rightarrow \{\pi^*(\cdot), m^*(\cdot)\}$, with either $\pi^*(\cdot) = \pi(\cdot)$ or $m^*(\cdot) = m(\cdot)$ **but not** necessarily both. Assume the same pointwise convergence conditions and rates $(\delta_{n,\pi}, \xi_{n,m})$ for $\{\widehat{\pi}(\cdot), \widehat{m}(\cdot)\}$ as in (2), but now with $\{\pi(\cdot), m(\cdot)\}$ therein replaced by $\{\pi^*(\cdot), m^*(\cdot)\}$.

- Under some 'suitable' assumptions, we have: with high probability,

$$\|\mathbf{T}_{0,n}\|_\infty + \|\mathbf{T}_{\pi,n}\|_\infty + \|\mathbf{T}_{m,n}\|_\infty \lesssim \sqrt{\frac{\log d}{n}} \left\{1 + 1_{(\pi^*, m^*) \neq (\pi, m)}\right\}$$

and $\|\mathbf{R}_{\pi,m,n}\|_\infty \lesssim \left\{\delta_{n,\pi} 1_{(m^* \neq m)} + \xi_{n,m} 1_{(\pi^* \neq \pi)} + \delta_{n,\pi}\xi_{n,m}\right\} (\log n)$.

- Finally, let $\{\widehat{\pi}(\cdot), \widehat{m}(\cdot)\} \to \{\pi^*(\cdot), m^*(\cdot)\}$, with either $\pi^*(\cdot) = \pi(\cdot)$ or $m^*(\cdot) = m(\cdot)$ **but not** necessarily both. Assume the same pointwise convergence conditions and rates $(\delta_{n,\pi}, \xi_{n,m})$ for $\{\widehat{\pi}(\cdot), \widehat{m}(\cdot)\}$ as in (2), but now with $\{\pi(\cdot), m(\cdot)\}$ therein replaced by $\{\pi^*(\cdot), m^*(\cdot)\}$.

- Under some 'suitable' assumptions, we have: with high probability,

$$\|\mathbf{T}_{0,n}\|_\infty + \|\mathbf{T}_{\pi,n}\|_\infty + \|\mathbf{T}_{m,n}\|_\infty \lesssim \sqrt{\frac{\log d}{n}} \left\{1 + 1_{(\pi^*, m^*) \neq (\pi, m)}\right\}$$

and $\|\mathbf{R}_{\pi,m,n}\|_\infty \lesssim \left\{\delta_{n,\pi} 1_{(m^* \neq m)} + \xi_{n,m} 1_{(\pi^* \neq \pi)} + \delta_{n,\pi}\xi_{n,m}\right\} (\log n)$.

The 2^{nd} and/or 3^{rd} terms also contribute now to the rate $\sqrt{(\log d)/n}$. The 4^{th} term is $o(1)$ but **no longer ignorable** (and may be slower).

- Finally, let $\{\widehat{\pi}(\cdot), \widehat{m}(\cdot)\} \rightarrow \{\pi^*(\cdot), m^*(\cdot)\}$, with either $\pi^*(\cdot) = \pi(\cdot)$ or $m^*(\cdot) = m(\cdot)$ **but not** necessarily both. Assume the same pointwise convergence conditions and rates $(\delta_{n,\pi}, \xi_{n,m})$ for $\{\widehat{\pi}(\cdot), \widehat{m}(\cdot)\}$ as in (2), but now with $\{\pi(\cdot), m(\cdot)\}$ therein replaced by $\{\pi^*(\cdot), m^*(\cdot)\}$.

- Under some 'suitable' assumptions, we have: with high probability,

$$\|\mathbf{T}_{0,n}\|_\infty + \|\mathbf{T}_{\pi,n}\|_\infty + \|\mathbf{T}_{m,n}\|_\infty \lesssim \sqrt{\frac{\log d}{n}} \left\{1 + 1_{(\pi^*,m^*)\neq(\pi,m)}\right\}$$

and $\|\mathbf{R}_{\pi,m,n}\|_\infty \lesssim \left\{\delta_{n,\pi} 1_{(m^*\neq m)} + \xi_{n,m} 1_{(\pi^*\neq\pi)} + \delta_{n,\pi}\xi_{n,m}\right\} (\log n)$.

The 2^{nd} and/or 3^{rd} terms also contribute now to the rate $\sqrt{(\log d)/n}$. The 4^{th} term is $o(1)$ but **no longer ignorable** (and may be slower).

  - Regardless, this establishes **general** convergence rates and the DR property of $\widehat{\boldsymbol{\theta}}_{\text{DDR}}$ under possible misspecification of $\{\widehat{\pi}(\cdot), \widehat{m}(\cdot)\}$. For the 4^{th} term, sharper rates **need** a case-by-case analysis.

- **Note:** our theory holds generally for **any** choices of $\widehat{\pi}(\cdot)$ and $\widehat{m}(\cdot)$ under mild conditions (provided they are both 'correct' estimators).
  - Under misspecifications, consistency & general non-sharp rates are also established. Sharp rates **need** case-by-case analyses.
    - **Even** for mean (or ATE) estimation problem, this can be quite tricky in HD settings. See Smucler et al. (2019) for a detailed analysis.

- **Note:** our theory holds generally for **any** choices of $\widehat{\pi}(\cdot)$ and $\widehat{m}(\cdot)$ under mild conditions (provided they are both 'correct' estimators).
    - Under misspecifications, consistency & general non-sharp rates are also established. Sharp rates **need** case-by-case analyses.
        - **Even** for mean (or ATE) estimation problem, this can be quite tricky in HD settings. See Smucler et al. (2019) for a detailed analysis.
- Below we provide only **some** choices of $\widehat{\pi}(\cdot)$ and $\widehat{m}(\cdot)$ that may be used to implement our theory & methods for $\widehat{\boldsymbol{\theta}}_{\mathrm{DDR}}$. In general, one can use any reasonable method (including black box ML methods).
- Choices of $\widehat{\pi}(\cdot)$ and $\widehat{m}(\cdot)$: we consider estimators from two families.

- **Note:** our theory holds generally for **any** choices of $\widehat{\pi}(\cdot)$ and $\widehat{m}(\cdot)$ under mild conditions (provided they are both 'correct' estimators).
  - Under misspecifications, consistency & general non-sharp rates are also established. Sharp rates **need** case-by-case analyses.
    - **Even** for mean (or ATE) estimation problem, this can be quite tricky in HD settings. See Smucler et al. (2019) for a detailed analysis.
- Below we provide only **some** choices of $\widehat{\pi}(\cdot)$ and $\widehat{m}(\cdot)$ that may be used to implement our theory & methods for $\widehat{\boldsymbol{\theta}}_{\mathrm{DDR}}$. In general, one can use any reasonable method (including black box ML methods).
- Choices of $\widehat{\pi}(\cdot)$ and $\widehat{m}(\cdot)$: we consider estimators from two families.
  - Parametric and 'extended' parametric families (series estimators).
  - Semi-parametric single index families.

Choices of $\widehat{\pi}(\cdot)$: 'Extended' Parametric Families (Series Estimators)

- If $\pi(\cdot)$ is known, we set $\widehat{\pi}(\cdot) := \pi(\cdot)$. Otherwise, we estimate $\pi(\cdot)$ via two (class of) choices of $\widehat{\pi}(\cdot)$ (each assumed to be 'correct').

- If $\pi(\cdot)$ is known, we set $\widehat{\pi}(\cdot) := \pi(\cdot)$. Otherwise, we estimate $\pi(\cdot)$ via two (class of) choices of $\widehat{\pi}(\cdot)$ (each assumed to be 'correct').
  - 'Extended' parametric family: $\pi(\mathbf{x}) = g\{\boldsymbol{\alpha}'\boldsymbol{\Psi}(\mathbf{X})\}$, where $g(\cdot) \in [0,1]$ is a **known** function [e.g. $g_{\text{expit}}(u) := \exp(u)/\{1+\exp(u)\}$], $\boldsymbol{\Psi}(\mathbf{X}) := \{\psi_k(\mathbf{X})\}_{k=1}^{K}$ is **any** set of K basis functions (with $K \gg n$ possibly), and $\boldsymbol{\alpha} \in \mathbb{R}^K$ is an unknown (sparse) parameter vector.

- If $\pi(\cdot)$ is known, we set $\widehat{\pi}(\cdot) := \pi(\cdot)$. Otherwise, we estimate $\pi(\cdot)$ via two (class of) choices of $\widehat{\pi}(\cdot)$ (each assumed to be 'correct').
  - 'Extended' parametric family: $\pi(\mathbf{x}) = g\{\boldsymbol{\alpha}'\boldsymbol{\Psi}(\mathbf{X})\}$, where $g(\cdot) \in [0,1]$ is a **known** function [e.g. $g_{\text{expit}}(u) := \exp(u)/\{1+\exp(u)\}$], $\boldsymbol{\Psi}(\mathbf{X}) := \{\psi_k(\mathbf{X})\}_{k=1}^{K}$ is **any** set of K basis functions (with $K \gg n$ possibly), and $\boldsymbol{\alpha} \in \mathbb{R}^K$ is an unknown (sparse) parameter vector.
    - Example: $\boldsymbol{\Psi}(\mathbf{X})$ may correspond to the **polynomial bases** of $\mathbf{X}$ upto any fixed degree k. Note: the special case of linear bases ($k = 1$) **includes** all standard parametric regression models. Further, the case of $\pi(\cdot) =$ constant (but unknown) i.e. MCAR is **also included.**

- If $\pi(\cdot)$ is known, we set $\widehat{\pi}(\cdot) := \pi(\cdot)$. Otherwise, we estimate $\pi(\cdot)$ via two (class of) choices of $\widehat{\pi}(\cdot)$ (each assumed to be 'correct').
  - 'Extended' parametric family: $\pi(\mathbf{x}) = g\{\boldsymbol{\alpha}'\boldsymbol{\Psi}(\mathbf{X})\}$, where $g(\cdot) \in [0,1]$ is a **known** function [e.g. $g_{\text{expit}}(u) := \exp(u)/\{1+\exp(u)\}$], $\boldsymbol{\Psi}(\mathbf{X}) := \{\psi_k(\mathbf{X})\}_{k=1}^{K}$ is **any** set of K basis functions (with $K \gg n$ possibly), and $\boldsymbol{\alpha} \in \mathbb{R}^K$ is an unknown (sparse) parameter vector.
    - Example: $\boldsymbol{\Psi}(\mathbf{X})$ may correspond to the **polynomial bases** of $\mathbf{X}$ upto any fixed degree k. Note: the special case of linear bases ($k=1$) **includes** all standard parametric regression models. Further, the case of $\pi(\cdot) =$ constant (but unknown) i.e. MCAR is **also included.**
    - Estimator: we set $\widehat{\pi}(\mathbf{X}) = g\{\widehat{\boldsymbol{\alpha}}'\boldsymbol{\Psi}(\mathbf{X})\}$, where $\widehat{\boldsymbol{\alpha}}$ denotes any suitable estimator (possibly penalized) of $\boldsymbol{\alpha}$ based on $\mathcal{T}_n := \{T_i, \mathbf{X}_i\}_{i=1}^n$.

- If $\pi(\cdot)$ is known, we set $\widehat{\pi}(\cdot) := \pi(\cdot)$. Otherwise, we estimate $\pi(\cdot)$ via two (class of) choices of $\widehat{\pi}(\cdot)$ (each assumed to be 'correct').
  - 'Extended' parametric family: $\pi(\mathbf{x}) = g\{\boldsymbol{\alpha}'\boldsymbol{\Psi}(\mathbf{X})\}$, where $g(\cdot) \in [0,1]$ is a **known** function [e.g. $g_{\text{expit}}(u) := \exp(u)/\{1+\exp(u)\}$], $\boldsymbol{\Psi}(\mathbf{X}) := \{\psi_k(\mathbf{X})\}_{k=1}^{K}$ is **any** set of K basis functions (with $K \gg n$ possibly), and $\boldsymbol{\alpha} \in \mathbb{R}^K$ is an unknown (sparse) parameter vector.
    - Example: $\boldsymbol{\Psi}(\mathbf{X})$ may correspond to the **polynomial bases** of $\mathbf{X}$ upto any fixed degree k. Note: the special case of linear bases ($k = 1$) **includes** all standard parametric regression models. Further, the case of $\pi(\cdot) =$ constant (but unknown) i.e. MCAR is **also included.**
    - Estimator: we set $\widehat{\pi}(\mathbf{X}) = g\{\widehat{\boldsymbol{\alpha}}'\boldsymbol{\Psi}(\mathbf{X})\}$, where $\widehat{\boldsymbol{\alpha}}$ denotes any suitable estimator (possibly penalized) of $\boldsymbol{\alpha}$ based on $\mathcal{T}_n := \{T_i, \mathbf{X}_i\}_{i=1}^n$.
    - **Example** of $\widehat{\boldsymbol{\alpha}}$: when $g(\cdot) = g_{\text{expit}}(\cdot)$, $\widehat{\boldsymbol{\alpha}}$ may be obtained based on a standard L_1-penalized logistic regression of $\{T_i$ vs. $\boldsymbol{\Psi}(\mathbf{X}_i)\}_{i=1}^n$.

- Semi-parametric single index family: $\pi(\mathbf{X}) = g(\boldsymbol{\alpha}'\mathbf{X})$, where $g(\cdot) \in (0,1)$ is **unknown** and $\boldsymbol{\alpha} \in \mathbb{R}^p$ is a (sparse) unknown parameter (identifiable **only** upto scalar multiples, hence set $\|\boldsymbol{\alpha}\|_2 = 1$ wlog).

- Semi-parametric single index family: $\pi(\mathbf{X}) = g(\boldsymbol{\alpha}'\mathbf{X})$, where $g(\cdot) \in (0,1)$ is **unknown** and $\boldsymbol{\alpha} \in \mathbb{R}^p$ is a (sparse) unknown parameter (identifiable **only** upto scalar multiples, hence set $\|\boldsymbol{\alpha}\|_2 = 1$ wlog).
  - Given an estimator $\widehat{\boldsymbol{\alpha}}$ of $\boldsymbol{\alpha}$, we estimate $\pi(\mathbf{X}) \equiv \mathbb{E}(T \mid \boldsymbol{\alpha}'\mathbf{X})$ as:

$$\widehat{\pi}(\mathbf{x}) \;\equiv\; \widehat{\pi}(\widehat{\boldsymbol{\alpha}}, \mathbf{x}) \;:=\; \frac{\frac{1}{nh}\sum_{i=1}^{n} T_i K\left\{\widehat{\boldsymbol{\alpha}}'(\mathbf{X}_i - \mathbf{x})/h\right\}}{\frac{1}{nh}\sum_{i=1}^{n} K\left\{\widehat{\boldsymbol{\alpha}}'(\mathbf{X}_i - \mathbf{x})/h\right\}},$$

  where $K(\cdot)$ denotes any standard (2^{nd} order) kernel function and $h = h_n > 0$ denotes the bandwidth sequence with $h = o(1)$.

- Semi-parametric single index family: $\pi(\mathbf{X}) = g(\boldsymbol{\alpha}'\mathbf{X})$, where $g(\cdot) \in (0,1)$ is **unknown** and $\boldsymbol{\alpha} \in \mathbb{R}^p$ is a (sparse) unknown parameter (identifiable **only** upto scalar multiples, hence set $\|\boldsymbol{\alpha}\|_2 = 1$ wlog).
  - Given an estimator $\widehat{\boldsymbol{\alpha}}$ of $\boldsymbol{\alpha}$, we estimate $\pi(\mathbf{X}) \equiv \mathbb{E}(T \mid \boldsymbol{\alpha}'\mathbf{X})$ as:

$$\widehat{\pi}(\mathbf{x}) \;\equiv\; \widehat{\pi}(\widehat{\boldsymbol{\alpha}}, \mathbf{x}) \;:=\; \frac{\frac{1}{nh}\sum_{i=1}^{n} T_i K\left\{\widehat{\boldsymbol{\alpha}}'(\mathbf{X}_i - \mathbf{x})/h\right\}}{\frac{1}{nh}\sum_{i=1}^{n} K\left\{\widehat{\boldsymbol{\alpha}}'(\mathbf{X}_i - \mathbf{x})/h\right\}},$$

    where $K(\cdot)$ denotes any standard (2^{nd} order) kernel function and $h = h_n > 0$ denotes the bandwidth sequence with $h = o(1)$.
  - Obtaining $\widehat{\boldsymbol{\alpha}}$: In general, **any approach** (if available) from (high dimensional) single index model literature can be used. But if $\mathbf{X}$ is elliptically symmetric, then $\widehat{\boldsymbol{\alpha}}$ may be obtained as simply as a standard L_1-penalized logistic regression of $\{T_i \text{ vs. } \mathbf{X}_i\}_{i=1}^{n}$.

- 'Extended' parametric family: $m(\mathbf{x}) = g\{\gamma' \mathbf{\Psi}(\mathbf{X})\}$, where $g(\cdot)$ is a **known** 'link' function [e.g. 'canonical' links: identity, expit or exp], $\mathbf{\Psi}(\mathbf{X}) := \{\psi_k(\mathbf{X})\}_{k=1}^K$ is **any** set of K basis functions (with $K \gg n$ possibly), and $\gamma \in \mathbb{R}^K$ is an unknown (sparse) parameter vector.

- 'Extended' parametric family: $m(\mathbf{x}) = g\{\gamma' \boldsymbol{\Psi}(\mathbf{X})\}$, where $g(\cdot)$ is a **known** 'link' function [e.g. 'canonical' links: identity, expit or exp], $\boldsymbol{\Psi}(\mathbf{X}) := \{\psi_k(\mathbf{X})\}_{k=1}^{K}$ is **any** set of K basis functions (with $K \gg n$ possibly), and $\gamma \in \mathbb{R}^K$ is an unknown (sparse) parameter vector.
  - Example: $\boldsymbol{\Psi}(\mathbf{X})$ may correspond to the **polynomial bases** of $\mathbf{X}$ upto any fixed degree k. Note: the special case of linear bases ($k = 1$) **includes** all standard parametric regression models.

- 'Extended' parametric family: $m(\mathbf{x}) = g\{\gamma' \boldsymbol{\Psi}(\mathbf{X})\}$, where $g(\cdot)$ is a **known** 'link' function [e.g. 'canonical' links: identity, expit or exp], $\boldsymbol{\Psi}(\mathbf{X}) := \{\psi_k(\mathbf{X})\}_{k=1}^K$ is **any** set of K basis functions (with $K \gg n$ possibly), and $\gamma \in \mathbb{R}^K$ is an unknown (sparse) parameter vector.
  - Example: $\boldsymbol{\Psi}(\mathbf{X})$ may correspond to the **polynomial bases** of $\mathbf{X}$ upto any fixed degree k. Note: the special case of linear bases ($k = 1$) **includes** all standard parametric regression models.
  - Estimator: we set $\widehat{m}(\mathbf{X}) = g\{\widehat{\gamma}' \boldsymbol{\Psi}(\mathbf{X})\}$, where $\widehat{\gamma}$ denotes any suitable estimator (possibly penalized) of γ based on the data subset of **'complete cases':** $\mathcal{D}_n^{(c)} := \{(Y_i, \mathbf{X}_i) \mid T_i = 1\}_{i=1}^n$.

- 'Extended' parametric family: $m(\mathbf{x}) = g\{\gamma' \boldsymbol{\Psi}(\mathbf{X})\}$, where $g(\cdot)$ is a **known** 'link' function [e.g. 'canonical' links: identity, expit or exp], $\boldsymbol{\Psi}(\mathbf{X}) := \{\psi_k(\mathbf{X})\}_{k=1}^K$ is **any** set of K basis functions (with $K \gg n$ possibly), and $\gamma \in \mathbb{R}^K$ is an unknown (sparse) parameter vector.
  - Example: $\boldsymbol{\Psi}(\mathbf{X})$ may correspond to the **polynomial bases** of $\mathbf{X}$ upto any fixed degree k. Note: the special case of linear bases ($k = 1$) **includes** all standard parametric regression models.
  - Estimator: we set $\widehat{m}(\mathbf{X}) = g\{\widehat{\gamma}' \boldsymbol{\Psi}(\mathbf{X})\}$, where $\widehat{\gamma}$ denotes any suitable estimator (possibly penalized) of γ based on the data subset of **'complete cases':** $\mathcal{D}_n^{(c)} := \{(Y_i, \mathbf{X}_i) \mid T_i = 1\}_{i=1}^n$.
  - **Example** of $\widehat{\gamma}$: when $g(\cdot) :=$ any 'canonical' link function, $\widehat{\gamma}$ may be simply obtained based on the respective usual L_1-penalized 'canonical' link based regression (e.g. linear, logistic or poisson) of $\{(Y_i \text{ vs. } \mathbf{X}_i) \mid T_i = 1\}_{i=1}^n$ from the 'complete case' data $\mathcal{D}_n^{(c)}$.

- Semi-parametric single index family: $m(\mathbf{X}) = g(\gamma'\mathbf{X})$, where $g(\cdot)$ is an **unknown** 'link' and $\gamma \in \mathbb{R}^p$ is a (sparse) unknown parameter (identifiable **only** upto scalar multiples, hence set $\|\gamma\|_2 = 1$ wlog).

- Semi-parametric single index family: $m(\mathbf{X}) = g(\gamma'\mathbf{X})$, where $g(\cdot)$ is an **unknown** 'link' and $\gamma \in \mathbb{R}^p$ is a (sparse) unknown parameter (identifiable **only** upto scalar multiples, hence set $\|\gamma\|_2 = 1$ wlog).
- Given an estimator $\widehat{\gamma}$ of γ, we estimate $m(\mathbf{X}) \equiv \mathbb{E}(Y \mid \gamma'\mathbf{X}, T)$ as:

$$\widehat{m}(\mathbf{x}) \; \equiv \; \widehat{m}(\widehat{\gamma}, \mathbf{x}) \; := \; \frac{\frac{1}{nh}\sum_{i=1}^{n} T_i Y_i \, K\left\{\widehat{\gamma}'(\mathbf{X}_i - \mathbf{x})/h\right\}}{\frac{1}{nh}\sum_{i=1}^{n} T_i \, K\left\{\widehat{\gamma}'(\mathbf{X}_i - \mathbf{x})/h\right\}},$$

where $K(\cdot)$ denotes any standard (2^{nd} order) kernel function, and $h = h_n > 0$ denotes the bandwidth sequence with $h = o(1)$.

- Semi-parametric single index family: $m(\mathbf{X}) = g(\gamma'\mathbf{X})$, where $g(\cdot)$ is an **unknown** 'link' and $\gamma \in \mathbb{R}^p$ is a (sparse) unknown parameter (identifiable **only** upto scalar multiples, hence set $\|\gamma\|_2 = 1$ wlog).

- Given an estimator $\widehat{\gamma}$ of γ, we estimate $m(\mathbf{X}) \equiv \mathbb{E}(Y \mid \gamma'\mathbf{X}, T)$ as:

$$\widehat{m}(\mathbf{x}) \;\equiv\; \widehat{m}(\widehat{\gamma}, \mathbf{x}) \;:=\; \frac{\frac{1}{nh}\sum_{i=1}^{n} T_i Y_i \, K\left\{\widehat{\gamma}'(\mathbf{X}_i - \mathbf{x})/h\right\}}{\frac{1}{nh}\sum_{i=1}^{n} T_i \, K\left\{\widehat{\gamma}'(\mathbf{X}_i - \mathbf{x})/h\right\}},$$

  where $K(\cdot)$ denotes any standard (2^{nd} order) kernel function, and $h = h_n > 0$ denotes the bandwidth sequence with $h = o(1)$.

- Obtaining $\widehat{\gamma}$: In general, **any approach** (if available) from HD SIM literature can be used on the complete case data subset $\mathcal{D}_n^{(c)}$.

  - If $\mathbf{X}$ is elliptically symmetric and $Y = f(\gamma'\mathbf{X}; \epsilon)$ with f **unknown** and $\epsilon \perp\!\!\!\perp (T, \mathbf{X})$, then $\widehat{\gamma}$ may be obtained as L_1-penalized IPW estimator $\widehat{\boldsymbol{\theta}}_{\mathrm{IPW}}$ for any 'canonical' link based regression problem.

- For either choices of $\widehat{\pi}(\cdot)$, assume that the ingredient estimator $\widehat{\boldsymbol{\alpha}}$ satisfies: $\|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\|_1 \lesssim_{\mathbb{P}} a_n$ for some $a_n = o(1)$. Then, under suitable smoothness and tail assumptions, with high probability (w.h.p.),

- For either choices of $\widehat{\pi}(\cdot)$, assume that the ingredient estimator $\widehat{\boldsymbol{\alpha}}$ satisfies: $\|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\|_1 \lesssim_{\mathbb{P}} a_n$ for some $a_n = o(1)$. Then, under suitable smoothness and tail assumptions, with high probability (w.h.p.),

$$|\widehat{\pi}(\mathbf{x}) - \pi(\mathbf{x})| \; \lesssim \; a_n \; = \; o(1), \quad \text{for any fixed } \mathbf{x} \in \mathcal{X}, \quad \text{(for method 1).}$$

- For either choices of $\widehat{\pi}(\cdot)$, assume that the ingredient estimator $\widehat{\boldsymbol{\alpha}}$ satisfies: $\|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\|_1 \lesssim_{\mathbb{P}} a_n$ for some $a_n = o(1)$. Then, under suitable smoothness and tail assumptions, with high probability (w.h.p.),

$$|\widehat{\pi}(\mathbf{x}) - \pi(\mathbf{x})| \lesssim a_n = o(1), \quad \text{for any fixed } \mathbf{x} \in \mathcal{X}, \quad \text{(for method 1).}$$

- For method 2 (SIM), assume that $h = o(1), \log(np)/(nh) = o(1)$ and $(a_n/h)\sqrt{\log p} = o(1)$. Then, under some suitable smoothness and tail assumptions, we have: with high probability, for any fixed $\mathbf{x} \in \mathcal{X}$,

$$|\widehat{\pi}(\mathbf{x}) - \pi(\mathbf{x})| \lesssim \left(h^2 + \frac{1}{\sqrt{nh}}\right) + \left(a_n + \frac{\log(np)}{nh} + \frac{a_n^2}{h^2}\right) = o(1).$$

- For either choices of $\widehat{\pi}(\cdot)$, assume that the ingredient estimator $\widehat{\boldsymbol{\alpha}}$ satisfies: $\|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\|_1 \lesssim_{\mathbb{P}} a_n$ for some $a_n = o(1)$. Then, under suitable smoothness and tail assumptions, with high probability (w.h.p.),

$$|\widehat{\pi}(\mathbf{x}) - \pi(\mathbf{x})| \ \lesssim \ a_n \ = \ o(1), \quad \text{for any fixed } \mathbf{x} \in \mathcal{X}, \quad \text{(for method 1)}.$$

- For method 2 (SIM), assume that $h = o(1), \log(np)/(nh) = o(1)$ and $(a_n/h)\sqrt{\log p} = o(1)$. Then, under some suitable smoothness and tail assumptions, we have: with high probability, for any fixed $\mathbf{x} \in \mathcal{X}$,

$$|\widehat{\pi}(\mathbf{x}) - \pi(\mathbf{x})| \ \lesssim \ \left(h^2 + \frac{1}{\sqrt{nh}}\right) + \left(a_n + \frac{\log(np)}{nh} + \frac{a_n^2}{h^2}\right) \ = \ o(1).$$

- Usually, we expect the L_1 error rate of $\widehat{\boldsymbol{\alpha}}$ to be $a_n = s_{\boldsymbol{\alpha}}\sqrt{(\log d_*)/n}$ where $s_\alpha := \|\boldsymbol{\alpha}\|_0$ and $d_* = K$ or p (depending on the method).

- For either choices of $\widehat{m}(\cdot)$, assume that the ingredient estimator $\widehat{\gamma}$ satisfies: $\|\widehat{\gamma} - \gamma\|_1 \lesssim_{\mathbb{P}} b_n$ for some $b_n = o(1)$. Then, under suitable smoothness and tail assumptions, we have: with high probability,

- For either choices of $\widehat{m}(\cdot)$, assume that the ingredient estimator $\widehat{\gamma}$ satisfies: $\|\widehat{\gamma} - \gamma\|_1 \lesssim_{\mathbb{P}} b_n$ for some $b_n = o(1)$. Then, under suitable smoothness and tail assumptions, we have: with high probability,

$$|\widehat{m}(\mathbf{x}) - m(\mathbf{x})| \lesssim b_n = o(1) \quad \text{for any fixed } \mathbf{x} \in \mathcal{X} \quad \text{(for method 1).}$$

- For either choices of $\widehat{m}(\cdot)$, assume that the ingredient estimator $\widehat{\gamma}$ satisfies: $\|\widehat{\gamma} - \gamma\|_1 \lesssim_{\mathbb{P}} b_n$ for some $b_n = o(1)$. Then, under suitable smoothness and tail assumptions, we have: with high probability,

$$|\widehat{m}(\mathbf{x}) - m(\mathbf{x})| \lesssim b_n = o(1) \quad \text{for any fixed } \mathbf{x} \in \mathcal{X} \quad \text{(for method 1).}$$

- For method 2 (SIM), assume that $h = o(1), \log(np)/(nh) = o(1)$ and $(a_n/h)\sqrt{\log p} = o(1)$. Then, under some suitable smoothness and tail assumptions, we have: with high probability, for any fixed $\mathbf{x} \in \mathcal{X}$,

$$|\widehat{m}(\mathbf{x}) - m(\mathbf{x})| \lesssim \left(h^2 + \frac{1}{\sqrt{nh}}\right) + \left(b_n + \frac{\log(np)}{nh} + \frac{b_n^2}{h^2}\right) = o(1).$$

- For either choices of $\widehat{m}(\cdot)$, assume that the ingredient estimator $\widehat{\gamma}$ satisfies: $\|\widehat{\gamma} - \gamma\|_1 \lesssim_{\mathbb{P}} b_n$ for some $b_n = o(1)$. Then, under suitable smoothness and tail assumptions, we have: with high probability,

$$|\widehat{m}(\mathbf{x}) - m(\mathbf{x})| \lesssim b_n = o(1) \quad \text{for any fixed } \mathbf{x} \in \mathcal{X} \quad \text{(for method 1).}$$

- For method 2 (SIM), assume that $h = o(1), \log(np)/(nh) = o(1)$ and $(a_n/h)\sqrt{\log p} = o(1)$. Then, under some suitable smoothness and tail assumptions, we have: with high probability, for any fixed $\mathbf{x} \in \mathcal{X}$,

$$|\widehat{m}(\mathbf{x}) - m(\mathbf{x})| \lesssim \left(h^2 + \frac{1}{\sqrt{nh}}\right) + \left(b_n + \frac{\log(np)}{nh} + \frac{b_n^2}{h^2}\right) = o(1).$$

- We typically expect the L_1 error rate of $\widehat{\gamma}$ to be $b_n = s_\gamma\sqrt{(\log d_*)/n}$ where $s_\gamma := \|\boldsymbol{\alpha}\|_0$ and $d_* = K$ or p (depending on the method).

Simulation Studies: The Setup

- Basic parameters: $n = 1000$, $p = 50$ or 500 and $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_p)$.
- Three data generating processes (DGPs) for $Y|\mathbf{X}$ and $T|\mathbf{X}$ as follows:

Simulation Studies: The Setup

- Basic parameters: $n = 1000$, $p = 50$ or 500 and $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_p)$.
- Three data generating processes (DGPs) for $Y|\mathbf{X}$ and $T|\mathbf{X}$ as follows:

1. "Linear-Linear" DGP:

$$\begin{aligned} Y &= \gamma_0 + \boldsymbol{\gamma}'\mathbf{X} + \varepsilon, \quad \varepsilon|\mathbf{X} \sim \mathcal{N}(0,1). \\ \mathrm{logit}\{\pi(\mathbf{X})\} \equiv \mathrm{logit}\{\mathbb{E}(T|\mathbf{X})\} &= \alpha_0 + \boldsymbol{\alpha}'\mathbf{X}. \end{aligned}$$

- Basic parameters: $n = 1000$, $p = 50$ or 500 and $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_p)$.
- Three data generating processes (DGPs) for $Y|\mathbf{X}$ and $T|\mathbf{X}$ as follows:

1. "Linear-Linear" DGP:

$$\begin{aligned} Y &= \gamma_0 + \boldsymbol{\gamma}'\mathbf{X} + \varepsilon, \quad \varepsilon|\mathbf{X} \sim \mathcal{N}(0,1). \\ \text{logit}\{\pi(\mathbf{X})\} \equiv \text{logit}\{\mathbb{E}(T|\mathbf{X})\} &= \alpha_0 + \boldsymbol{\alpha}'\mathbf{X}. \end{aligned}$$

2. "Quad-Quad" DGP:

$$Y = \gamma_0 + \boldsymbol{\gamma}'\mathbf{X} + \sum_{j=1}^{p} \gamma_j^* \mathbf{X}_j^2 + \varepsilon, \quad \varepsilon|\mathbf{X} \sim \mathcal{N}(0,1).$$

$$\text{logit}\{\pi(\mathbf{X})\} \equiv \text{logit}\{\mathbb{E}(T|\mathbf{X})\} = \alpha_0 + \boldsymbol{\alpha}'\mathbf{X}_i + \sum_{j=1}^{p} \boldsymbol{\alpha}_j^* \mathbf{X}_{ij}^2.$$

- Basic parameters: $n = 1000$, $p = 50$ or 500 and $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_p)$.
- Three data generating processes (DGPs) for $Y|\mathbf{X}$ and $T|\mathbf{X}$ as follows:

1. "Linear-Linear" DGP:

$$\begin{aligned} Y &= \gamma_0 + \boldsymbol{\gamma}'\mathbf{X} + \varepsilon, \quad \varepsilon|\mathbf{X} \sim \mathcal{N}(0,1). \\ \mathrm{logit}\{\pi(\mathbf{X})\} \equiv \mathrm{logit}\{\mathbb{E}(T|\mathbf{X})\} &= \alpha_0 + \boldsymbol{\alpha}'\mathbf{X}. \end{aligned}$$

2. "Quad-Quad" DGP:

$$Y = \gamma_0 + \boldsymbol{\gamma}'\mathbf{X} + \sum_{j=1}^{p} \gamma_j^* \mathbf{X}_j^2 + \varepsilon, \quad \varepsilon|\mathbf{X} \sim \mathcal{N}(0,1).$$

$$\mathrm{logit}\{\pi(\mathbf{X})\} \equiv \mathrm{logit}\{\mathbb{E}(T|\mathbf{X})\} = \alpha_0 + \boldsymbol{\alpha}'\mathbf{X}_i + \sum_{j=1}^{p} \boldsymbol{\alpha}_j^* \mathbf{X}_{ij}^2.$$

3. "SIM-SIM" DGP:

$$\begin{aligned} Y &= \gamma_0 + \boldsymbol{\gamma}'\mathbf{X} + c_Y(\boldsymbol{\gamma}'\mathbf{X})^2 + \varepsilon, \quad \varepsilon|\mathbf{X} \sim \mathcal{N}(0,1). \\ \mathrm{logit}\{\pi(\mathbf{X})\} \equiv \mathrm{logit}\{\mathbb{E}(T|\mathbf{X})\} &= \alpha_0 + \boldsymbol{\alpha}'\mathbf{X} + c_T(\boldsymbol{\alpha}'\mathbf{X})^2. \end{aligned}$$

- Choices of the parameters:
  1. Covariance matrix $\mathbf{\Sigma}_p$ (for today): $\mathbf{\Sigma}_p = I_p$ (identity matrix).
  2. We set $c_T = 0.2$, $c_Y = 0.3$ and $\gamma_0 = 1$, $\alpha_0 = 0.5$.
  3. When $p = 50$, $\boldsymbol{\alpha} = 1/\sqrt{5}(1, -1, 0.5, -0.5, 0.5, 0, \cdots, 0)$ with $\|\boldsymbol{\alpha}\|_0 = 5$, $\boldsymbol{\gamma} = (1, 1, 1, -1, -1, 0.5, 0.5, -0.5, -0.5, -0.5, 0, \cdots, 0)$ with $\|\boldsymbol{\gamma}\|_0 = 10$, $\boldsymbol{\alpha}^* = (0.25, -0.25, 0, \cdots, 0)$ and $\boldsymbol{\gamma}^* = (1, -1, 0.5, 0.5, -0.5, 0, \cdots, 0)$.

Simulation Settings: Choice of Parameters

- Choices of the parameters:

  1. Covariance matrix $\boldsymbol{\Sigma}_p$ (for today): $\boldsymbol{\Sigma}_p = I_p$ (identity matrix).
  2. We set $c_T = 0.2$, $c_Y = 0.3$ and $\gamma_0 = 1$, $\alpha_0 = 0.5$.
  3. When $p = 50$, $\boldsymbol{\alpha} = 1/\sqrt{5}(1, -1, 0.5, -0.5, 0.5, 0, \cdots, 0)$ with $\|\boldsymbol{\alpha}\|_0 = 5$, $\boldsymbol{\gamma} = (1, 1, 1, -1, -1, 0.5, 0.5, -0.5, -0.5, -0.5, 0, \cdots, 0)$ with $\|\boldsymbol{\gamma}\|_0 = 10$, $\boldsymbol{\alpha}^* = (0.25, -0.25, 0, \cdots, 0)$ and $\boldsymbol{\gamma}^* = (1, -1, 0.5, 0.5, -0.5, 0, \cdots, 0)$.
  4. When $p = 500$, $\|\boldsymbol{\alpha}\|_0 = 10$ and $\boldsymbol{\alpha}$ consists of three 1s, two -1s, two 0.5s and three -0.5s normalized by $1/\sqrt{10}$, while $\|\boldsymbol{\gamma}\|_0 = 15$ and $\boldsymbol{\gamma}$ consists of three 1s, two -1s, five 0.5s, five -0.5s, two 0.25s and three -0.25s. Further, we set $\boldsymbol{\alpha}^* = (0.25, 0.25, -0.25, -0.25, 0, \cdots, 0)$ and $\boldsymbol{\gamma}^* = (1, -1, 0.5, 0.5, -0.5, 0, \cdots, 0)$.

- Choices of the parameters:

  1. Covariance matrix $\mathbf{\Sigma}_p$ (for today): $\mathbf{\Sigma}_p = I_p$ (identity matrix).
  2. We set $c_T = 0.2$, $c_Y = 0.3$ and $\gamma_0 = 1$, $\alpha_0 = 0.5$.
  3. When $p = 50$, $\boldsymbol{\alpha} = 1/\sqrt{5}(1, -1, 0.5, -0.5, 0.5, 0, \cdots, 0)$ with $\|\boldsymbol{\alpha}\|_0 = 5$, $\boldsymbol{\gamma} = (1, 1, 1, -1, -1, 0.5, 0.5, -0.5, -0.5, -0.5, 0, \cdots, 0)$ with $\|\boldsymbol{\gamma}\|_0 = 10$, $\boldsymbol{\alpha}^* = (0.25, -0.25, 0, \cdots, 0)$ and $\boldsymbol{\gamma}^* = (1, -1, 0.5, 0.5, -0.5, 0, \cdots, 0)$.
  4. When $p = 500$, $\|\boldsymbol{\alpha}\|_0 = 10$ and $\boldsymbol{\alpha}$ consists of three 1s, two -1s, two 0.5s and three -0.5s normalized by $1/\sqrt{10}$, while $\|\boldsymbol{\gamma}\|_0 = 15$ and $\boldsymbol{\gamma}$ consists of three 1s, two -1s, five 0.5s, five -0.5s, two 0.25s and three -0.25s. Further, we set $\boldsymbol{\alpha}^* = (0.25, 0.25, -0.25, -0.25, 0, \cdots, 0)$ and $\boldsymbol{\gamma}^* = (1, -1, 0.5, 0.5, -0.5, 0, \cdots, 0)$.

- $\mathbb{K} = 2$ fold cross-fitting used; all simulation settings replicated 500 times.

- $\widehat{\boldsymbol{\Omega}}$ obtained as $\widehat{\boldsymbol{\Sigma}}^{-1}$ for $p = 50$ and using the nodewise Lasso for $p = 500$.

Simulation Settings: Estimators Implemented

- Obtain the DDR estimator $\widehat{\boldsymbol{\theta}}_{\mathrm{DDR}}$ for linear regression: $\boldsymbol{\theta}_0 = \boldsymbol{\Sigma}^{-1}\mathbb{E}(\mathbf{X}Y)$.

Simulation Settings: Estimators Implemented

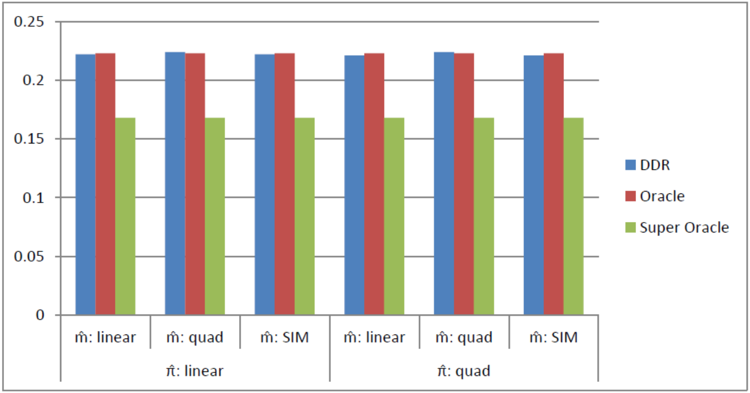
- Obtain the DDR estimator $\widehat{\boldsymbol{\theta}}_{\mathrm{DDR}}$ for linear regression: $\boldsymbol{\theta}_0 = \boldsymbol{\Sigma}^{-1}\mathbb{E}(\mathbf{X}Y)$.
- Two choices of the working nuisance models for $\pi(\mathbf{X})$ to obtain $\widehat{\pi}(\mathbf{X})$:
  1. Linear: L_1 penalized logistic-linear regression.
  2. Quad: L_1 penalized logistic-linear regression with quadratic terms.

- Obtain the DDR estimator $\widehat{\boldsymbol{\theta}}_{\mathrm{DDR}}$ for linear regression: $\boldsymbol{\theta}_0 = \boldsymbol{\Sigma}^{-1}\mathbb{E}(\mathbf{X}Y)$.
- Two choices of the working nuisance models for $\pi(\mathbf{X})$ to obtain $\widehat{\pi}(\mathbf{X})$:
  1. Linear: L_1 penalized logistic-linear regression.
  2. Quad: L_1 penalized logistic-linear regression with quadratic terms.
- Three choices of the working nuisance models for $m(\mathbf{X})$ to obtain $\widehat{m}(\mathbf{X})$:
  1. Linear: L_1 penalized linear regression.
  2. Quad: L_1 penalized linear regression with quadratic terms.
  3. SIM: Single index model (with index parameter estimated via IPW Lasso)

Simulation Settings: Estimators Implemented

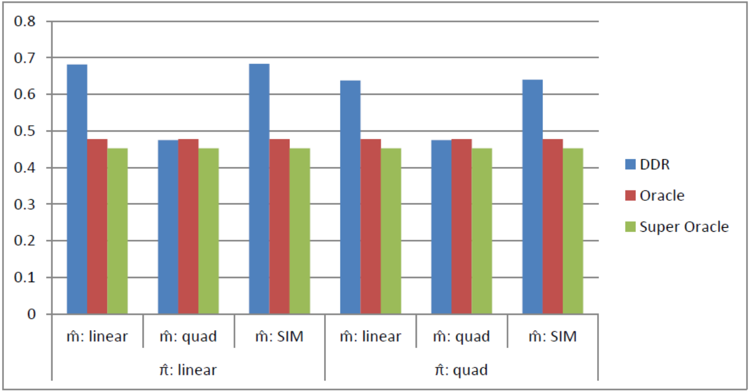
- Obtain the DDR estimator $\widehat{\boldsymbol{\theta}}_{\mathrm{DDR}}$ for linear regression: $\boldsymbol{\theta}_0 = \boldsymbol{\Sigma}^{-1}\mathbb{E}(\mathbf{X}Y)$.
- Two choices of the working nuisance models for $\pi(\mathbf{X})$ to obtain $\widehat{\pi}(\mathbf{X})$:
  1. Linear: L_1 penalized logistic-linear regression.
  2. Quad: L_1 penalized logistic-linear regression with quadratic terms.
- Three choices of the working nuisance models for $m(\mathbf{X})$ to obtain $\widehat{m}(\mathbf{X})$:
  1. Linear: L_1 penalized linear regression.
  2. Quad: L_1 penalized linear regression with quadratic terms.
  3. SIM: Single index model (with index parameter estimated via IPW Lasso)
- Estimators used for comparison:
  1. $\widehat{\boldsymbol{\theta}}_{orac}$ (Oracle): obtained assuming both $\pi(\cdot)$ and $m(\cdot)$ are known.
  2. $\widehat{\boldsymbol{\theta}}_{full}$ (Super oracle): obtained assuming a full dataset is observed.
- Criteria: L_2 errors for estimation and coverage probability for inference.

$p = 50$, DGP: Linear-Linear.

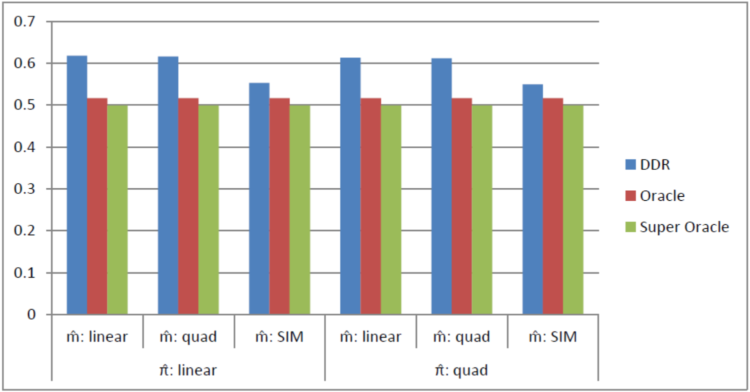


Simulation Results: L_2 Error Comparison ($p = 50$) - II

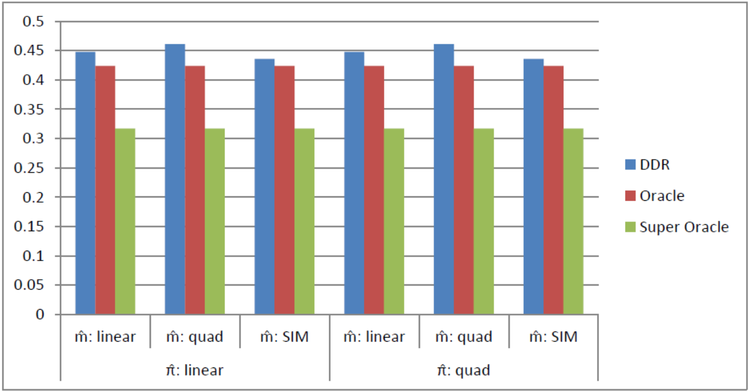
$p = 50$, DGP: Quad-Quad.



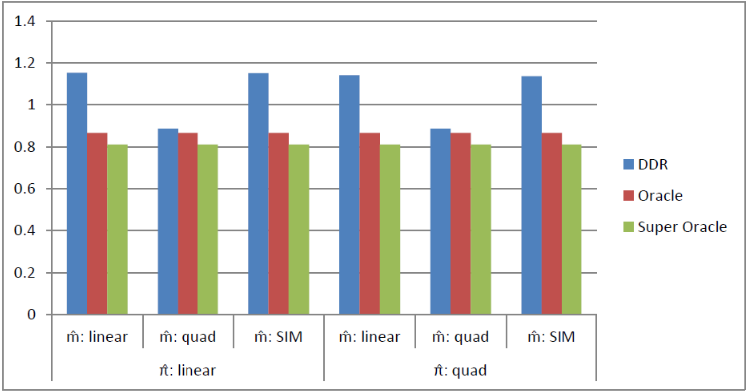
$p = 50$, DGP: SIM-SIM.



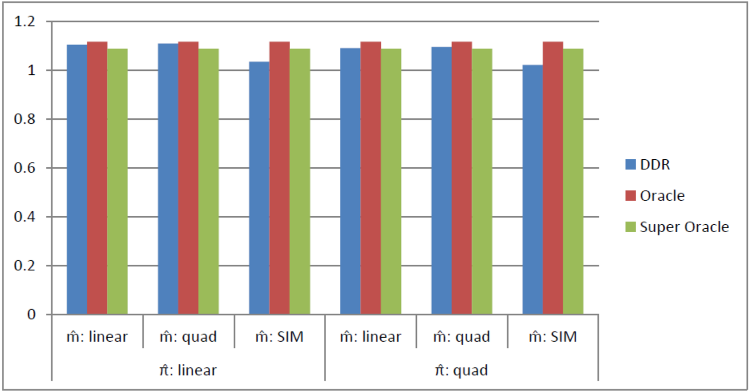
$p = 500$, DGP: Linear-Linear.



$p = 500$, DGP: Quad-Quad.



$p = 500$, DGP: SIM-SIM.



Coverage probability (covg. prob.) of the DDR estimator:

DGP: Linear-Linear.

Coverage probability (covg. prob.) of the DDR estimator:

DGP: Linear-Linear.

1. When $p = 50$:

	$\widehat{m}$: linear	$\widehat{m}$: quad	$\widehat{m}$: SIM	$\widehat{m}$: linear	$\widehat{m}$: quad	$\widehat{m}$: SIM
	Average Covg. Prob. (zero coeffs.)			Average Covg. Prob. (non-zero coeffs.)		
$\widehat{\pi}$: logit	0.94 (0.01)	0.94 (0.01)	0.95 (0.01)	0.94 (0.01)	0.94 (0.01)	0.93 (0.01)
$\widehat{\pi}$: quad	0.94 (0.01)	0.95 (0.01)	0.95 (0.01)	0.94 (0.01)	0.94 (0.01)	0.94 (0.01)

2. When $p = 500$:

	$\widehat{m}$: linear	$\widehat{m}$: quad	$\widehat{m}$: SIM	$\widehat{m}$: linear	$\widehat{m}$: quad	$\widehat{m}$: SIM
	Average Covg. Prob. (zero coeffs.)			Average Covg. Prob. (non-zero coeffs.)		
$\widehat{\pi}$: logit	0.94 (0.01)	0.94 (0.01)	0.94 (0.01)	0.92 (0.01)	0.91 (0.02)	0.92 (0.01)
$\widehat{\pi}$: quad	0.94 (0.01)	0.94 (0.01)	0.94 (0.01)	0.91 (0.02)	0.91 (0.02)	0.92 (0.01)

Coverage probability (covg. prob.) of the DDR estimator:

DGP: Quad-Quad.

Coverage probability (covg. prob.) of the DDR estimator:

DGP: Quad-Quad.

1. When $p = 50$:

	$\widehat{m}$: linear	$\widehat{m}$: quad	$\widehat{m}$: SIM	$\widehat{m}$: linear	$\widehat{m}$: quad	$\widehat{m}$: SIM
	Average Covg. Prob. (zero coeffs.)			Average Covg. Prob. (non-zero coeffs.)		
$\widehat{\pi}$: logit	0.94 (0.01)	0.94 (0.01)	0.95 (0.01)	0.88 (0.16)	0.94 (0.01)	0.88 (0.16)
$\widehat{\pi}$: quad	0.95 (0.01)	0.94 (0.01)	0.95 (0.01)	0.89 (0.12)	0.94 (0.01)	0.89 (0.12)

2. When $p = 500$:

	$\widehat{m}$: linear	$\widehat{m}$: quad	$\widehat{m}$: SIM	$\widehat{m}$: linear	$\widehat{m}$: quad	$\widehat{m}$: SIM
	Average Covg. Prob. (zero coeffs.)			Average Covg. Prob. (non-zero coeffs.)		
$\widehat{\pi}$: logit	0.95 (0.01)	0.94 (0.01)	0.95 (0.01)	0.91 (0.03)	0.92 (0.01)	0.91 (0.05)
$\widehat{\pi}$: quad	0.95 (0.01)	0.94 (0.01)	0.95 (0.01)	0.91 (0.03)	0.92 (0.01)	0.91 (0.04)

Coverage probability (covg. prob.) of the DDR estimator:

DGP: SIM-SIM.

Coverage probability (covg. prob.) of the DDR estimator:

DGP: SIM-SIM.

1. When $p = 50$:

	$\widehat{m}$: linear	$\widehat{m}$: quad	$\widehat{m}$: SIM	$\widehat{m}$: linear	$\widehat{m}$: quad	$\widehat{m}$: SIM
	Average Covg. Prob. (zero coeffs.)			Average Covg. Prob. (non-zero coeffs.)		
$\widehat{\pi}$: logit	0.94 (0.01)	0.95 (0.01)	0.95 (0.01)	0.94 (0.01)	0.94 (0.01)	0.94 (0.01)
$\widehat{\pi}$: quad	0.94 (0.01)	0.95 (0.01)	0.95 (0.01)	0.94 (0.01)	0.94 (0.01)	0.94 (0.01)

2. When $p = 500$:

	$\widehat{m}$: linear	$\widehat{m}$:quad	$\widehat{m}$: SIM	$\widehat{m}$: linear	$\widehat{m}$:quad	$\widehat{m}$: SIM
	Average Covg. Prob. (zero coeffs.)			Average Covg. Prob. (non-zero coeffs.)		
$\widehat{\pi}$: logit	0.94 (0.01)	0.95 (0.01)	0.95 (0.01)	0.87 (0.05)	0.88 (0.04)	0.93 (0.02)
$\widehat{\pi}$: quad	0.94 (0.01)	0.95 (0.01)	0.95 (0.01)	0.87 (0.05)	0.87 (0.05)	0.93 (0.02)

Consider $n = 50000$ and $p = 50$. In addition, also consider the complete case estimator $\widehat{\boldsymbol{\theta}}_{cc}$ (obtained by using only the data with $T_i = 1$).

DGP: Quad-Quad ($p = 50$)

Consider $n = 50000$ and $p = 50$. In addition, also consider the complete case estimator $\widehat{\boldsymbol{\theta}}_{cc}$ (obtained by using only the data with $T_i = 1$).

DGP: Quad-Quad ($p = 50$)

L_2 Error Comparison:

model		$\widehat{\boldsymbol{\theta}}_{\text{DDR}}$	$\widehat{\boldsymbol{\theta}}_{orac}$	$\widehat{\boldsymbol{\theta}}_{full}$	$\widehat{\boldsymbol{\theta}}_{cc}$
$\widehat{m}$: linear	$\widehat{\pi}$: logit	0.460 (0.026)	0.072 (0.011)	0.069 (0.01)	0.528 (0.021)
	$\widehat{\pi}$: quad	0.204 (0.137)	0.072 (0.011)	0.069 (0.01)	0.528 (0.021)
$\widehat{m}$: quad	$\widehat{\pi}$: logit	0.071 (0.010)	0.072 (0.011)	0.069 (0.01)	0.528 (0.021)
	$\widehat{\pi}$: quad	0.072 (0.011)	0.072 (0.011)	0.069 (0.01)	0.528 (0.021)
$\widehat{m}$: SIM	$\widehat{\pi}$: logit	0.323 (0.019)	0.072 (0.011)	0.069 (0.01)	0.528 (0.021)
	$\widehat{\pi}$: quad	0.172 (0.078)	0.072 (0.011)	0.069 (0.01)	0.528 (0.021)

Inference:

	$\widehat{m}$: linear	$\widehat{m}$:quad	$\widehat{m}$: SIM	$\widehat{m}$: linear	$\widehat{m}$:quad	$\widehat{m}$: SIM
	Average Covg. Prob. (zero coeffs.)			Average Covg. Prob. (non-zero coeffs.)		
$\widehat{\pi}$: logit	0.94 (0.03)	0.94 (0.03)	0.94 (0.03)	0.68 (0.39)	0.93 (0.03)	0.80 (0.19)
$\widehat{\pi}$: quad	0.96 (0.02)	0.94 (0.03)	0.95 (0.02)	0.96 (0.02)	0.94 (0.02)	0.95 (0.02)

Consider $n = 50000$ and $p = 500$. In addition, also consider the complete case estimator $\widehat{\boldsymbol{\theta}}_{cc}$ (obtained by using only the data with $T_i = 1$).

DGP: Quad-Quad ($p = 500$)

Consider $n = 50000$ and $p = 500$. In addition, also consider the complete case estimator $\widehat{\boldsymbol{\theta}}_{cc}$ (obtained by using only the data with $T_i = 1$).

DGP: Quad-Quad ($p = 500$)

L_2 Error Comparison:

model		$\widehat{\boldsymbol{\theta}}_{\mathrm{DDR}}$	$\widehat{\boldsymbol{\theta}}_{orac}$	$\widehat{\boldsymbol{\theta}}_{full}$	$\widehat{\boldsymbol{\theta}}_{cc}$
$\widehat{m}$: linear	$\widehat{\pi}$: logit	0.297 (0.017)	0.178 (0.009)	0.173 (0.007)	0.325 (0.018)
	$\widehat{\pi}$: quad	0.282 (0.113)	0.178 (0.009)	0.173 (0.007)	0.325 (0.018)
$\widehat{m}$: quad	$\widehat{\pi}$: logit	0.177 (0.008)	0.178 (0.009)	0.173 (0.007)	0.325 (0.018)
	$\widehat{\pi}$: quad	0.180 (0.010)	0.178 (0.009)	0.173 (0.007)	0.325 (0.018)
$\widehat{m}$: SIM	$\widehat{\pi}$: logit	0.407 (0.022)	0.178 (0.009)	0.173 (0.007)	0.325 (0.018)
	$\widehat{\pi}$: quad	0.294 (0.045)	0.178 (0.009)	0.173 (0.007)	0.325 (0.018)

Inference:

	$\widehat{m}$: linear	$\widehat{m}$:quad	$\widehat{m}$: SIM	$\widehat{m}$: linear	$\widehat{m}$:quad	$\widehat{m}$: SIM
	Average Covg. Prob. (zero coeffs.)			Average Covg. Prob. (non-zero coeffs.)		
$\widehat{\pi}$: logit	0.95 (0.02)	0.95 (0.02)	0.95 (0.02)	0.78 (0.32)	0.94 (0.02)	0.75 (0.38)
$\widehat{\pi}$: quad	0.95 (0.02)	0.95 (0.02)	0.95 (0.02)	0.94 (0.04)	0.94 (0.02)	0.88 (0.12)

References I

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973.

Belloni, A., Chernozhukov, V., Fernández-Val, I., and Hansen, C. (2017). Program evaluation and causal inference with high-dimensional data. *Econometrica*, 85(1):233–298.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68.

Chernozhukov, V., Demirer, M., Duflo, E., and Fernandez-Val, I. (2017). Generic machine learning inference on heterogenous treatment effects in randomized experiments. *ArXiv preprint arXiv:1712.04802*.

Chernozhukov, V. and Semenova, V. (2017). Simultaneous inference for best linear predictor of the conditional average treatment effect and other structural functions. *ArXiv preprint arXiv:1702.06240v2*.

Farrell, M. H. (2015). Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics*, 189(1):1–23.

Li, K.-C. and Duan, N. (1989). Regression analysis under link violation. *The Annals of Statistics*, 17(3):1009–1052.

References II

Loh, P.-L. and Wainwright, M. J. (2012). High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *The Annals of Statistics*, 40(3):1637.

Loh, P.-L. and Wainwright, M. J. (2015). Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *Journal of Machine Learning Research*, 16:559–616.

Negahban, S. N., Ravikumar, P., Wainwright, M. J., and Yu, B. (2012). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866.

Smucler, E., Rotnitzky, A., and Robins, J. M. (2019). A unifying approach for doubly-robust l_1 regularized estimation of causal contrasts. *ArXiv preprint arXiv:1904.03737v1*.

Tsiatis, A. (2007). *Semiparametric Theory and Missing Data*. Springer Science & Business Media.

van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal cnfidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202.

Van der Vaart, A. W. (2000). *Asymptotic Statistics*, volume 3. Cambridge University Press.

Thank You!