目次

第 1 章

研究展望

ネタ帳

(1) 二重ロバスト推定と Double Machine Learning.

(2) 高次の影響関数への推定手法 (Robins が開拓) は因果推論ではとても有望視されている．Gateaux 微分ではだめでも，Malliavin 微分で理論を創れないか？これは robust 統計学の塗り替えに当たる試みになってしまう．

(3) Gaussian 近似．鈴木太慈先生と今泉先生の分野に近い．

1 で売れて，2 を温めるしかない．

Fisher 情報量は KL 情報量の二次形式．基本的に解析学と同じ発展経路をたどるんだろうな．

頑健性が反脆弱性になったら機械学習なんじゃないのか？ミスから学ぶ．

影響関数が存在するようなパラメータ付けを正則というらしい．となると統計モデルの解析学はここからじゃないと始まらないじゃないか！

1.1 頑健性についてのノート

1.1.1 解題

用語法 1.1.1. 古典的な統計手法は，仮定された正規分布よりも尾が重い標本分布に対してとても感度が高く，すなわち外れ値に対して脆弱である．これを distributional distortion に対する distributional robustness という．一部では resistant 統計量ともいう．そして，robust の語を，モデル設定や推定量についての仮定への違反に対して使うために控える．だが，通常 robustness といったとき，distributional robustness を指す．外れ値を，経験分布とデルタ分布との凸結合によって表して，これの Gateaux 片側微分係数によって定まる影響関数で，頑健性の定義とするのがひとつの流儀である．これを B-ロバスト統計量とよんでいる．

(1) B-ロバスト統計量は，バイアスについて頑健である．

例 1.1.2. 確率分布の中心的傾向 (central tendency) の指標として，算術平均は脆弱であるが，中央値は頑健である．M-推定量は普通頑健な方である．

1.1.2 情報量規準

そのモデルがはじき出すパラメータの推定量が，どの程度真の分布に近いかを KL 情報量によって測るとしたら，平均対数尤度の最大化を考えれば良い．しかし真の分布は未知であるから経験測度を代わりに用いることになるが，ここで同じデータを 2 度使うことになるので，これが原因で正のバイアスが生じる．[†1] これをなんとかする手法が情報量規準である．

小西-北川による一般化情報量規準にも影響関数が登場する．さらに石黒による拡張情報量規準はブートストラップ法によって対数尤度のバイアスを推定する手法で，標本分割法のように，バイアスの明示的な表現をしない計算機科学的な解決だと理解できる．

[†1] 3 ページ目第 2.3 節の 3 4 行目．https://www.ism.ac.jp/editsec/toukei/pdf/47-2-375.pdf

1.2　Inference for Semiparametric Models

直近 25 年のノンパラ，セミパラの発展 (契機は当然計算資源の増加) の概観をし，5 つの open question を提示する 2001 論文．そのあとの Robins and Rotnitzky のコメントが頑健性について言及していて Chernozhukov に引用されている．その他も多くの研究者がコメントを寄せているが，Robins and Rotnitzky は非常に大きなセミパラモデルにおいて，どのようにしてパラメータ θ を推定するかの問題を第 6 として挙げていて，部分的な回答として 2 重頑健性を議論している．

1.2.1　Introduction

計算資源が増加してノンパラが使えるようになったが，その認知容易性の低さはセミパラが補うことになった．特に生存解析の Cox モデル，計量経済学の index model.

1.2.2　Robins and Rotnitzky : Motivation

記法 1.2.1. Y を結果，$R \in 2$ を割当，V を共変量とすると，疫学研究ではよく $500 \leqslant n \leqslant 2000$，$V \subset \mathbb{R}^m$ $(50 \leqslant m \leqslant 100)$ となり，Y, V は連続として良い．V が高次元な連続量であるので，nonparametric smoothing は次元の呪いで使えず，また従来の層別などの共変量処理も連続量なので好ましくない．

例 1.2.2 (outcome regression model)**.** 典型的に歳用される統計モデルは，

$$E[Y|R,V] = \beta_0 + \beta^\top V + \theta R$$

なる線型結果回帰 (OR: outcome regression) で，最小二乗法を行う．θ が平均処置効果であり，$\beta_0 + \beta^\top V$ が共変量による影響である．線型ではなくセミパラメトリックな回帰モデル

$$E[Y|R,V] = \omega(V) + \theta R$$

は，ω に対するノンパラメトリックな smoothing が出来ないので，採用不可能である．すると当然，

(1) ω が線型から遠く，非線形項の R との相関が大きく，
(2) highly predictive of Y

であるとき，misspecification によるバイアスが大きくなってしまう．さらにこの検出が難しい．

(1) 推定された回帰関数 $\widehat{\beta}^\top_{OLS,0} + \widehat{\beta}^\top_{OLS} V + \widehat{\theta}_{OR} R$ は Y をうまく予測し，
(2) global lack of fit test の出力も小さい

ことがありえる．また，従来の次数削減技法も採用できない．GAM(General Additive Model) モデルは V の要素の間の相関を無視する．

例 1.2.3 (propensity score model)**.** 1983 に開発された confounder control の手法は，$P := P[R=1|V]$ とおく．$\mathrm{logit}(x) = \log\left(\dfrac{x}{1-x}\right), \mathrm{expit}(x) = \dfrac{1}{1+e^{-x}}$ とする．ノンパラメトリックなロジスティック回帰

$$\mathrm{logit} P[R=1|V] = \gamma(V)$$

は実行できない．そこで，線型ロジスティック回帰

$$\mathrm{logit} P[R=1|V] = \alpha_0 + \alpha^\top V$$

を，最尤法により $\widehat{P} := \mathrm{expit}\left(\widehat{\alpha_0} + \widehat{\alpha}^\top V\right)$ により推定する．

これにより共変量調整をして得られる推定量 $\widehat{\theta}_P$ は θ の，モデル

$$E[Y|R,V] = \beta_0 + \theta R + \zeta \widehat{P}$$

による最小二乗法推定量に等価になる (Robins 2000).

要諦 1.2.4 (doubly robust / doubly-protected methods)**.** 前者は結果回帰のモデルがミスるとバイアスが生じ，後者は処置回帰のモデルがミスるとバイアスが生じる．このままでは単に，「どの部分をパラメトリックにするか」の違いでしかない．1 つの打開策は，統計的実験列の極限に対して，片方のモデルが正しければバイアスが生じなくなるようなセミパラメトリックモデルの構成である．

例 1.2.5. 前述の線型結果回帰モデルに，項 $\zeta\widehat{P}$ を加えて得るモデル

$$E[Y|R,V] = \beta_0 + \beta^\top V + \theta R + \zeta\widehat{P}$$

に対する OLS 推定量 $\widehat{\theta}_{DR}$ は二重に頑健である．

要諦 1.2.6. これは，モデルの設定に成功した $\widehat{\theta}_{OR}$ より効率性は劣る．どちらを取るかである．さらに，$\widehat{\theta}_P, \widehat{\theta}_{OR}, \widehat{\theta}_{DR}$ の 3 つを比較することで，ある程度 goodness of fit について何が起こっているかの指標を得られる．

問題 1.2.7. 二重に頑健な推定量が存在するのはいつか？

(1) 線型回帰モデルに $\widehat{P}$ 項を加えると DR 推定量が得られることは見た．
(2) 線型ロジスティックモデル $\mathrm{logit}E[Y|R,V] = \beta_0 + \beta^\top V + \theta R$ ($Y \sim$ Bernoulli) には DR 推定量は存在しない．
(3) 対数線型モデル $\log E[Y|R,V] = \beta_0 + \beta^\top V + \theta R$ (Y : count variable) には DR 推定量は存在するが，$\widehat{P}$ 項をモデルに加えることによって構成される訳ではない．

1.2.3　The Formal Problem and Doubly Robust Estimating Functions

以降，Bickel らが打ち立てたセミパラ理論により，問題の定式化と部分的な回答を行う．

記法 1.2.8. モデル $M(\mathcal{R}) := (P_\rho)_{\rho\in\mathcal{R}}$ 上の関数 $\theta : \mathcal{R} \to \mathbb{R}^p$ を，i.i.d. 標本 $X_1, \cdots, X_n$ から推定することを考える．$\mathcal{R}$ は次元の呪いのために θ の直接の推定が難しいとする．というのも，次の 3 条件を仮定する．

(1) θ の $\sqrt{n}$-一致性を持つ推定量 $\widehat{\theta}$ の分散は任意の $\rho \in \mathcal{R}$ について有界．だが，どの推定量も一様に一致性を持つことも，一様漸近正規性を持つこともない．
(2) 次は成り立たない：ある推定量 $\widehat{\theta}$ が存在して，$\exists_{\alpha>0}\ \forall_{\rho\in\mathcal{R}}\ |\widehat{\theta} - \theta| = O(n^\alpha)$ $(n \to \infty)$.
(3) どの $\rho \in \mathcal{R}$ についても，θ の正則な漸近線型推定量は存在しない．

議論 1.2.9 (dimension reduction)**.** 次の次元削減を行って，モデルをセミパラメトリックにすることとなる．分解 $\mathcal{R} = \mathcal{K} \times \Gamma$ について ($\kappa \in \mathcal{K}, \gamma \in \Gamma$ は独立)，サブモデル $\mathcal{K}' \subset \mathcal{K}, \Gamma' \subset \Gamma$ を定め，$M(\mathcal{K} \times \Gamma') \cup M(\mathcal{K}' \times \Gamma)$ を代わりに考える．このとき，このモデルの推定量は，パラメータ付け $\rho = (\kappa, \gamma)$ について 2 重に頑健であるという．2 回チャンスがあり，$\mathcal{K} \times \Gamma'$ または $\mathcal{K}' \times \Gamma$ の中に真値があれば一致性が得られる．

1.2.3.1　一点集合の場合

$\mathcal{K}', \Gamma'$ が一点集合 κ, γ であるとき．

定義 1.2.10 (doubly robust estimating function)**.** $U(\theta, \kappa, \gamma) := u(X, \theta, \kappa, \gamma)$ は $\theta(\kappa, \gamma)$ のモデル $M(\mathcal{K} \times \Gamma)$ 下でのパラメータ付け (κ, γ) に関する 2 重に頑健な推定量であるとは，モデル $M(\kappa \times \Gamma) \cup M(\mathcal{K} \times \gamma)$ における不偏推定量であることをいう．すなわち，次の 2 条件が成り立つことをいう：

(1) $\forall_{(\kappa,\gamma),(\kappa^*,\gamma^*)}\ E_{\kappa^*,\gamma^*}[U(\theta(\kappa^*,\gamma^*),\kappa^*,\gamma)] = E_{\kappa^*,\gamma^*}[U(\theta(\kappa^*,\gamma^*),\kappa,\gamma^*)] = 0.$
(2) $\left.\dfrac{\partial E_{\kappa^*,\gamma^*}[U(\theta,\kappa,\gamma^*)]}{\partial\theta}\right|_{\theta=\theta(\kappa^*,\gamma^*)}, \left.\dfrac{\partial E_{\kappa^*,\gamma^*}[U(\theta,\kappa^*,\gamma)]}{\partial\theta}\right|_{\theta=\theta(\kappa^*,\gamma^*)}$ は可逆．

記法 1.2.11. $M(\Psi_1 \times \Psi_2)$ をセミパラメトリックモデル，$\theta : \Psi_1 \times \Psi_2 \to \mathbb{R}^p$ を汎関数とする．

定義 1.2.12. $\mathcal{L}_2^0(\psi)$ で，パラメータ ψ が定める確率分布の下で平均 0 であるような確率ベクトル全体のなす Hilbert 空間

$\mathcal{L}_2^0(\psi) := \{f \in L^2(\mathcal{X}, \mathcal{F}, P_\psi; \mathbb{R}^p) \mid E[f] = 0\}$ を表す．内積は各確率ベクトルの分散共分散行列となっていることに注意．

(1)

1.3 Locally Robust Semiparametric Estimation

2 重に頑健な推定量が存在するための十分条件を議論する．1st step に対して頑健な GMM 推定量を定める局所頑健＝直交モーメント関数の標準的構成法．1st step というのは，モデル選択や正則化が産むバイアスを含むので，特に機械学習に応用先がある．

1.3.1 Introduction

この論文では，一般化モーメント法の推定関数＝モーメント関数を，局所頑健＝直交に構成する方法を示す．これが定める一般化モーメント推定量を，ここでは脱偏済み推定量 (debiased estimator) という．

> We show that such moment functions can be constructed by adding to identifying moment functions the non-parametric influence function from the effect of the first step on identifying moments.

1.3.2 Debiased GMM estimator

記法 1.3.1. $\theta \in \Theta \subset \mathbb{R}^p$ を母数，$\gamma : \Xi \to \mathbb{R}$ を未知関数（局外母数），$W : \Omega \to \mathcal{X}$ を観察されたデータ，$w \in \mathcal{X}$ をその実現値とする．真値 $\theta_0 \in \Theta, \gamma_0 \in \Xi \times \mathbb{R}$ に関して $E[g(W, \gamma_0, \theta_0)] = 0$ を満たす $g :\to \mathbb{R}^q$ を所与の関数とする．$\theta_0 \in \Theta$ は $E[g(W, \gamma_0, \theta)] = 0$ を満たす唯一の値であるという識別性条件（すなわち，θ_0 を，一般化モーメント $E[g(W, \gamma_0, \theta)]$ の値で識別できる．このことを"**identifying moment function** g"と呼んでいる）を仮定する．

議論 1.3.2. まず，$\hat{\gamma}$ を γ_0 の first-step 推定量としよう．すると，plug-in method として，観測値 w_i と推定値 $\hat{\gamma}$ を代入して $g(w_i, \gamma, \theta)$ を推定関数とし，経験測度について平均を取ると，標本モーメントの推定量 $\frac{\sum_{i=1}^n g(w_i, \hat{\gamma}, \theta)}{n}$ を得る．これの特定のノルム (論文では quadratic form と表現されてる) を最小にする θ を，"plug-in" GMM estimator とすれば良い．しかし，この推定量は，最初の $\hat{\gamma}$ を得たときのモデル選択，ひいては正則化法に多大な影響を受ける．[†2]

1.3.2.1 直交モーメント関数への準備

F をモデルの CDF とし，F_0 を真の累積分布関数とする．推定量 $\hat{\gamma}$ に対して，F の関数 $\gamma(F)$ が存在して，$F = F_0$ のときこれは推定量 $\hat{\gamma}$ の確率収束極限であるとする．

H を別の分布関数とし，凸結合 $F_\tau := (1 - \tau)F_0 + \tau H$ を考える．絶対連続性など理想的な状況が揃い，$E[g(W, \gamma(F_\tau), \theta)]$ の $\tau = 0$ における右 Gateaux 微分係数の微分 ϕ を影響関数という：

$$\frac{d}{d\tau} E[g(W, \gamma(F_\tau), \theta)] = \int \phi(w, \gamma_0, \alpha_0, \theta) H(dw)$$

どういう統計量 T についての影響関数かというと，一般化モーメント $\mu(F) := E[g(W, \gamma(F), \theta)]$ に関する影響関数である．

この影響関数 $\phi(w, \gamma, \alpha, \theta)$ を**ノンパラメトリック影響関数**といい，最初の推定量 γ が，標本モーメントの推定値 $\mu(F)$ に与える局所的な影響を記述しているとみなせる．

ただし，$E[\phi(W, \gamma_0, \alpha_0, \theta)] = 0$ を満たすとする．

定義 1.3.3 (orthogonal moment function)**.** 識別可能性を持つ一般化モーメント関数 (identifying moment function)g と，ノンパラメトリック影響関数（の微分）ϕ の和

$$\psi(W, \gamma, \alpha, \theta) := g(W, \gamma, \theta) + \phi(W, \gamma, \alpha, \theta)$$

[†2] 星野先生の本では，first-step は無限次元ではなく，傾向スコアに関するパラメトリックモデルだとして並行な議論をしている．時代は進んだ．

を直交モーメント関数という．

要諦 1.3.4. ϕ が，γ の影響を打ち消すから，ψ は頑健になる，ということか！条件 $E[\phi(W,\gamma_0,\alpha_0,\theta)]=0$ により，任意の経験測度について，$\widehat{\phi}_n \xrightarrow{p} 0$ が成り立つ．よって，ϕ の影響はこの意味ではなく，ただ $\widehat{\gamma}_l$ の first-order effect を打ち消すだけの役割を果たす．また $\widehat{\phi}$ の構成も，モデルの仮定に依らず，関数解析の言葉でノンパラメトリックに指定している．

定理 1.3.5.

(1) γ,α の推定量による ψ の期待値への影響の 1 次の項は消える．
(2) θ の推定量による ϕ の期待値への影響の 1 次の項は消える．

1.3.2.2　cross-fitting

"plug-in" method と違って，θ も第一段階で推定してしまうのか．
そして，このようなことをする理由は，平均を取るところの標本の選び方からバイアスを抜くため．ある種のリサンプリング法？

この ψ を用いて，debiased GMM estimator を構成するのだが，ここで **cross-fitting** を用いる．これは sample splitting の一種である．標本の添字集合 $[n]$ を L 個に分割し，$\widehat{\gamma}_l,\widehat{\alpha},\widehat{\theta}_l$ $(l\in[L])$ を，成分 $I_l\subset[n]$ に属する標本を使わずに算出した推定量とする．そして，脱偏済み標本モーメント関数 $\widehat{\psi}$ を

$$\widehat{g}(\theta)=\frac{1}{n}\sum_{l=1}^{L}\sum_{i\in I_l} g(W_i,\widehat{\gamma}_l,\theta),\quad \widehat{\phi}=\frac{1}{n}\sum_{l=1}^{L}\sum_{i\in I_l}\phi(W_i,\widehat{\gamma}_l,\widehat{\alpha}_l,\widehat{\theta}_l)$$

について

$$\widehat{\psi}(\theta)=\widehat{g}(\theta)+\widehat{\phi}$$

と定め，特定の正な半正定値行列 $\widehat{W}$ について，$\widehat{\theta}:=\arg\min_{\theta\in\Theta}\widehat{\psi}(\theta)^\top\widehat{W}\widehat{\psi}(\theta)$ として debiased GMM estimator $\widehat{\theta}$ を定義すれば良い．

注 1.3.6 (θ の第一段階推定量について)**.** $\widehat{\theta}_l$ についてだが，通常の GMM として

$$\widehat{\theta}_l:=\arg\min_{\theta\in\Theta}\widehat{g}_l(\theta)^\top\widehat{\Upsilon}_l\widehat{g}_l(\theta),\quad \widehat{g}_l(\theta)=\frac{1}{n-n_l}\sum_{l'\neq l}\sum_{i\in I_{l'}} g(W_i,\widehat{\gamma}_{l'},\theta)$$

] と定めても良い．ただし，$\widehat{\Upsilon}_l$ の構成には I_l に入っていない標本のみを使い，cross-fitting を取り入れる．このあとに debiased GMM を計算するセットを，何度か繰り返して最終的な debiased GMM としても良い（本当か？）

1.3.2.3　効率

こうして得た debiased GMM の効率は，以下の 3 つの要因のみのよる．

(1) モーメント関数 g の選び方
(2) 第一段階推定量 $\widehat{\gamma},\widehat{\alpha},\widehat{\theta}$
(3) 重み付け行列 $\widehat{W}$

(3) については別の議論で済んでいて，標準的な取り方 Ψ がある．

1.3.2.4　例 1 ：条件付き共分散

例 1.3.7. $W=(X,Y,Z)$ を観測データ，$\alpha_0(X):=E[Z|X]$ を共変量と割当の関係，$\gamma_0(X):=E[Y|X]$ を共変量と平均処置効果の関係とする．モーメント関数を $g:=\alpha_0\otimes\gamma_0$ と定めて，$\theta_0=E[Z\gamma_0(X)]=E[\alpha_0(X)\gamma_0(X)]=E[E[Z|X]\cdot E[Y|X]]$ と仮定する．このような設定は，条件付き共分散の平均 $E[\mathrm{Cov}(Z,Y|X)]=E[ZY]-\theta_0$ などに登場する．

このとき，

$$g(w,\gamma,\theta):=z\gamma(x)-\theta,\quad \phi(w,\gamma,\alpha)=\alpha(x)[y-\gamma(x)]$$

と定めると，

1.3.2.5　例 2 ：条件付き分位点の関数

1.3.3　例 3 ：

1.3.4　Neyman 直交性

Gateaux 微分における展開で一次項が消えていることをいう．どういう内積を仮定しているんだ？

定義 1.3.8. 未知関数 γ, α が，モーメント $\overline{\psi}(\gamma, \alpha, \theta) := E[\psi(W, \gamma, \alpha, \theta)]$ に 1 次の影響を及ぼさないことを，**Neyman 直交**するという．すなわち，F を変数 W の累積分布関数とするとき，$\widehat{\alpha} \xrightarrow{p} \alpha(F)$ と表すと，影響関数 $\phi(w, \gamma, \alpha, \theta)$ の識別可能性条件より，$0 = E_F[\phi(W, \gamma(F), \alpha(F), \theta)]$ が成り立つ．$F_\tau = (1-\tau)F_0 + \tau H$ を代入して $\tau = 0$ について微分することで，

$$\begin{aligned} 0 &= \frac{\partial}{\partial \tau} \int \phi(w, \gamma(F_\tau), \alpha(F_\tau), \theta)(-F_0 + H)(dw) \\ &= \int \phi(w, \gamma_0, \alpha_0, \theta) H(dw) + \frac{\partial}{\partial \tau} E[\phi(W, \gamma(F_\tau), \alpha(F_\tau), \theta)] \\ &= \frac{\partial}{\partial \tau} E[g(W, \gamma(F_\tau), \theta)] + \frac{\partial}{\partial \tau} E[\phi(W, \gamma(F_\tau), \alpha(F_\tau), \theta)] = \frac{\partial}{\partial \tau} \overline{\psi}(\gamma(F_\tau), \alpha(F_\tau), \theta) \end{aligned}$$

を得て，たしかに 1 次の微分係数がきえている．これは $(\gamma(F_\tau), \alpha(F_\tau))$ に関する方向微分だと思えるはず．

定理 1.3.9. 次の 3 条件が成り立つとき，Neyman 直交性を満たす．

(1) ノンパラメトリック影響関数の識別可能性（とは少しずれるらしい）：$E[\phi(W, \gamma_0, \alpha_0, \theta)] = 0$.
(2) 影響関数の微分可能性：ある $\overline{\tau} > 0$ が存在して，任意の $\tau \in [0, \overline{\tau})$ について $\int \phi(w, \gamma(F_\tau), \alpha(F_\tau), \theta) F_\tau(dw) = 0$.
(3) 連続性：$\int \phi(w, \gamma(F_\tau), \alpha(F_\tau), \theta) F_0(dw)$, $\int \phi(W, \gamma(F_\tau), \alpha(F_\tau), \theta) H(dw)$ はいずれも $\tau = 0$ で連続.

1.3.4.1　二重ロバスト性へむけて

定理 1.3.10 (α の十分条件)**.** α が次の 2 条件を満たすならば，$E[\phi(W, \gamma_0, \alpha, \theta)] = 0$. よって，$\overline{\psi} = E[\psi(W, \gamma_0, \alpha, \theta)] = 0$.

(1) F_α が存在して，$\alpha(F_\alpha) = \alpha$ かつ $\exists_{\overline{\tau}>0}\ \forall_{t \in \overline{\tau}}\ \gamma(F_\tau^\alpha) = \gamma_0$.
(2) $\dfrac{d}{d\tau} \int g(w, \gamma(F_\tau^\alpha), \theta) F_\alpha(dw) = \int \phi(w, \gamma_0, \alpha, \theta) F_0(dw)$.

定理 1.3.11 (γ の十分条件)**.** γ のノルム $\|-\|$ と，γ_0 を含む集合 Γ と，累積分布関数の集合 $\mathfrak{H}$ は次の 4 条件を満たすならば，$\forall_{\delta \in \Gamma}\ \overline{\psi}_\gamma(\delta, \alpha_0, \theta_0) = 0$

(1) $\alpha(F_\tau) = \alpha_0$ で，Neyman 直交性を満たす.
(2) $\overline{\psi(\gamma, \alpha_0, \theta_0)}$ は γ_0 にて，Γ に近接して Hadamard 微分可能で，Γ 上で定義された偏導関数 $\overline{\psi_\gamma}(\delta, \alpha_0, \theta_0)$ を持つ.
(3) $\gamma(F_\tau)$ は $\tau = 0$ にて Hadamard 微分可能.
(4) $\left\{ \dfrac{\partial \gamma(F_\tau)}{\partial \tau} \right\}_{H \in \mathfrak{H}}$ の閉包は Γ に等しい.

また，$\overline{\psi}(\gamma, \alpha_0, \theta_0)$ は γ_0 の近傍で二階連続微分可能で，近傍にて $\exists_{C \in \mathbb{R}}\ \|\overline{\psi}(\gamma, \alpha_0, \theta_0)\| \leqslant C\|\gamma - \gamma_0\|^2$.

1.3.5　Plug-in GMM との比較

1.3.6　α_0 の自動推定

1.3.7　二重ロバスト性

ノンパラメトリック影響関数を，identifying モーメント関数 g に加える，という手法で，二重ロバスト性を満たす ψ を構成できる条件を与える．

議論 1.3.12. 一般に，局外母数 γ を正しく推定することは困難なので，二重ロバスト性はもはや必須となりつつある．これは，モーメント関数 g に関するクラスで，ある二重ロバスト条件を満たすことをいう．いままでノンパラメトリック影響関数の mean zero condition は $E[\phi(W,\gamma_0,\alpha_0,\theta)]=0$ としてきたが，$E[\phi(W,\gamma,\alpha_0,\theta)]=E[\phi(W,\gamma_0,\alpha,\theta)]=0$ と強める．

1.3.7.1　二重ロバスト性の特徴付け

定義 1.3.13. Γ を γ の第一段階推定量のありえる値域となる位相線型空間とする．

$$\forall_{\gamma\in\Gamma}\ \forall_{\alpha,\theta}\quad 0=\overline{\psi}(\gamma,\alpha_0,\theta_0)=\overline{\psi}(\gamma_0,\alpha,\theta)$$

を満たすとき，モーメント $\overline{\psi}$ は二重に頑健であるという．

系 1.3.14. Γ を線型空間とする．$\overline{\psi}(\gamma,\alpha_0,\theta_0)$ が γ について線型で，定理 31.3.11 の条件が成り立つとき，二重に頑健である．

［証明］. 定理 31.3.11 の意味での Γ は γ の近傍であり，局所的な集合であったが，これが生成する線型空間の全体にまで，性質 $\overline{\psi}(\gamma,\alpha_0,\theta_0)=0$ が延長できるため． ■

定理 1.3.15 (Gateaux 微分のことばで述べ直す). Γ を線型空間とする．次の 2 条件は同値．

(1) $\psi(w,\gamma,\alpha,\theta)$ は二重に頑健である．
(2) $\overline{\psi}(\gamma,\alpha_0,\theta_0)$ は γ について affine で，$\forall_{\gamma\in\Gamma}\ \dfrac{\partial\overline{\psi}((1-\tau)\gamma_0+\tau\gamma,\alpha_0,\theta_0)}{\partial\tau}|_{\tau=0}=0.$

系 1.3.16. $g(W,\gamma,\theta_0)$ と $\phi(W,\gamma,\alpha_0,\theta_0)$ が γ について線型であるとき，これが定める $\overline{\psi}(\gamma,\alpha,\theta)$ は二重に頑健である．

1.3.7.2　第一推定量の条件付きモーメント条件下での二重頑健推定量の構成法

議論 1.3.17. 第一段階推定量 γ_0 は，ある線型汎関数 $\lambda(W,-):\Gamma\to\mathbb{R}$ について，$E[\lambda(W,\gamma_0)|X]=0$ を満たすとする．X は共変量や操作変数を想定すれば良い．

この下で 2SLS(two stage least squares) 推定量 $\hat{\gamma}$ を考えると，$\gamma(F):=\underset{\gamma\in\Gamma}{\arg\min}\,E_F[(E_F[\lambda(W,\gamma)|X])^2]$ として，$\phi(w,\gamma,\alpha,\theta)$ が存在するとき，ある $\alpha(x,\theta)$ について，

$$\phi(w,\gamma,\alpha,\theta)=\alpha(x,\theta)\lambda(W,\gamma)$$

が成り立つ．特に，ϕ は γ について線型である．よって，ψ の線形性は g の線形性による．

定理 1.3.18. 次の 2 条件は同値．

(1) $\psi(W,\gamma,\alpha,\theta)=g(W,\gamma,\theta)+\alpha(X)\lambda(W,\gamma)$ は二重に頑健である．
(2) $\forall_{\gamma\in\Gamma}\ E[g(W,\gamma,\theta_0)]=-E[\alpha_0(X)\lambda(W,\gamma)]$.

要諦 1.3.19. 有効スコアの場合に似ている関数形をもっているらしい．

例 1.3.20. γ を回帰関数，$\lambda(W,\gamma)=Y-\gamma(X)$ を誤差を表す線型汎関数，$m(w,\gamma)$ を線型汎関数とし，$\theta_0=E[m(W,\gamma_0)]$ の推定を目指す．モーメント関数を $g(w,\gamma,\theta):=m(w,\gamma)-\theta$ とおくと，これは識別可能な上に γ について線型で，λ も affine だから，定理 8 の前提条件を満たす．よって，

系 1.3.21. $\alpha_0(x) \in \mathcal{L}^2(\mathcal{X})$ が存在し，$\forall_{\gamma(X)\in\mathcal{L}^2(\mathcal{X})}\ E[m(W,\gamma)] = E[\alpha_0(X)\gamma(X)]$ ならば，$\psi(w,\gamma,\alpha,\theta) = m(w,\gamma) - \theta + \alpha(x)[y-\gamma(x)]$ は二重に頑健である．

1.3.7.3　局外母数が共変量の確率密度関数であるときの二重頑健推定量の構成

1.3.7.4　識別可能性

定理 1.3.22.

1.3.7.5　plug-in GMM 推定量の部分的頑健性

定義 1.3.23. モーメント関数 g が**部分的に頑健**であるとは，$\exists_{\overline{\gamma}\neq\gamma_0}\ E[g(W,\theta_0,\overline{\gamma})] = 0$ がなりたつことをいう．

1.3.8　漸近理論

1.4　Characterization of parameters with a mixed bias property

(1) 混合バイアス性を持ったパラメータのクラスに対する，セミパラメトリック有効なワンステップ推定量のバイアスは，2 つの局外関数の推定誤差の積の平均に等しい．
(2) ノンパラメトリックモデルにおける混合バイアス性を持つパラメータに対しては，2 つの局外関数の十分早いレートについて推定することに成功した場合に一致性と漸近正規性を持つような推定量 (レート二重に頑健な推定量 rate doubly robust estimator) が存在する．
(3) 混合バイアス性を持ったパラメータのクラスは，因果推論で近年提出された 2 つのクラスを真に含む．
(4) 混合バイアス性を持ったパラメータの表示と，影響関数の形を決定する．

1.4.1　Introduction

記法 1.4.1. P に独立に従う確率ベクトル $O : \Omega \to \mathcal{O}$ の列 $O^n : \Omega^n \to \mathcal{O}^n$ を考える．モデルを $\mathcal{M} := (P_\eta)_{\eta\in\eta}$ とし，汎関数 $\chi(\eta) \in \mathbb{R}$ の推定を試みる．η は Euclid 的でないとする．O は有限部分 Z を含み，その値域は $\mathbb{Z} \subset \mathbb{R}^d$ に収まるとする．

問題 1.4.2. 共変量 Z の未知関数 ($E[-|Z]$ など) を推定する必要があるパラメータ $\chi(\eta)$ の推定を考える．これは平均処置効果の例の一般化である．

議論 1.4.3 (debiased-GMM の動機：ワンステップ推定量)**.** 前の論文のように，plug-in 推定量 $\chi(\widehat{\eta})$ は一般に $\sqrt{n}$-一致性を持たない：$\chi(\widehat{\eta}) - \chi(\eta) \neq O_p(1/\sqrt{n})$．よって，漸近有効性が足りないから，バイアスを減らす余地がある．一つの解決法は，ワンステップ推定量（のセミパラメトリック化）である．χ^1_η をバイアスの 1 次項を打ち消すように選ぶことで，$\widehat{\chi} := \chi(\widehat{\eta}) + \mathbb{P}_n\chi^1_{\widehat{\eta}}$ をワンステップ推定量とする．χ^1_η の上手な選び方は，前の論文のように，影響関数とすることである．

議論 1.4.4 (影響関数を採用する動機)**.** 正規分布に収束させたい項 $\sqrt{n}(\widehat{\chi} - \chi(\eta))$ は

$$\sqrt{n}(\widehat{\chi} - \chi(\eta)) = \sqrt{n}(\chi(\widehat{\eta}) - \chi(\eta) + E_\eta(\chi^1_{\widehat{\eta}})) + \mathbb{G}_n(\chi^1_{\widehat{\eta}} - \chi^1_\eta) + \mathbb{G}(\chi^1_\eta)$$

と展開できる．

(a) 第 3 項 $\mathbb{G}_n(\chi^1_{\widehat{\eta}}) = \sqrt{n}\mathbb{P}_n(\chi^1_\eta - E_\eta(\chi^1_{\widehat{\eta}}))$ は正規分布に収束する．
(b) 第 2 項 $\mathbb{G}_n(\chi^1_{\widehat{\eta}} - \chi^1_\eta)$ は $o_p(1)$ である．一般にモデル $\mathcal{M}$ が大きくないとき (Donsker であるとき) はこれが示せ，一般の場合も cross-fitting によりこれが可能になる．$\mathcal{D}_n$ を複数に分割し，一部で推定量を構成し，残った部分でワンステップ推定量を構成する．これを繰り返し，全部の平均を最終的な推定量とする．
(c) 第 1 項が $o_p(1)$ ならば，$\sqrt{n}(\widehat{\chi} - \chi(\eta))$ は正規分布に確率収束する．

すなわち，$\chi(\widehat{\eta}) - \chi(\eta) + E_\eta(\chi^1_{\widehat{\eta}}) = o_p(n^{-1/2})$ が満たされれば良い．$E_\eta(\chi^1_{\widehat{\eta}})$ が，$\chi(\eta)$ の方向 $\widehat{\eta} - \eta$ への微分係数となっていれば良いわけである．こうして微分概念が必要になる．

定義 1.4.5 (正則パラメータ)**.** 影響関数が存在するようなパラメータ付けを正則という．

補題 1.4.6. $\mathcal{M}$ がノンパラメトリックであるとき，正則なパラメータ付け $\chi(\eta)$ は一意な影響関数を持つ．すなわち，任意の P におけるパラメトリックな部分モデルのスコアが生成する閉線型部分空間は $L_2(P)$ に同型である．

1.4.1.1　枠組み

因果推論における指数型分布族？
この mixed bias property ということばで，いままでの 2 つのクラス（Robins et al 2008 と Chernozhukov et al 2018）をまとめ上げた．モーメント関数の識別可能性条件などは，自動的に満たされる．

記法 1.4.7. $\mathcal{M}$ はノンパラメトリックとする：任意の $P \in \mathcal{M}$ におけるパラメトリックなサブモデルのスコアが生成する閉線形空間は $L_2(P)$ に等しい．$\chi(\eta)$ を正則なパラメータ付けとし，χ^1_η をその一意な影響関数とする．次の mixed bias property を満たすとする．

定義 1.4.8 (mixed bias property)**.** 任意のパラメータ $\eta \in \eta$ について，関数 $a(Z) := a(Z;\eta)$ と $b(Z) := b(Z;\eta)$ が存在し，また η に依らない関数 $S_{ab} := s_{ab}(O) : \mathcal{O} \to \mathbb{R}; o \mapsto s_{ab}(o)$ が存在して，

$$\forall_{\eta' \in \eta} \quad \chi(\eta') - \chi(\eta) + E_\eta(\chi^1_{\eta'}) = E_\eta[S_{ab}(a'(Z) - a(Z))(b'(Z) - b(Z))]$$

が成り立つ．ただし，$a'(Z) = a(Z;\eta'), b'(Z) = b(Z;\eta')$ と表した．

要諦 1.4.9. 特に，$\chi(\eta) + \chi^1_\eta$ は，a, b という 2 つの関数を通じてのみ η に依存しており，したがってワンステップ推定量 $\widehat{\chi(\eta)}$ は推定量 $\hat{a}, \hat{b}$ を通じてのみ $\hat{\eta}$ に依存する．

このことにより，$\int (\hat{a}(z) - a(z))^2 dP_\eta(z) = O_p(\gamma_{a,n})$ かつ $\int (\hat{b}(z) - b(z))^2 dP_\eta(z) = O_p(\gamma_{b,n})$ を満たすとき，$\chi(\hat{\eta}) - \chi(\eta) - E_\eta(\chi^1_{\hat{\eta}}) = O_p(\gamma_{a,n}\gamma_{b,n})$ が成り立つ．だから，cross-fitting が採用されたならば，$\tilde{\chi}$ は **rate double robustness property** を持つ．すなわち，$\gamma_{a,n} = o(1), \gamma_{b,n} = o(1), \gamma_{a,n}\gamma_{b,n} = o(n^{-1/2})$ が成り立つならば，$\sqrt{n}(\tilde{\chi} - \chi(\eta))$ は平均 0 の正規分布に収束する．

γ の収束レートは関数 a, b の複雑性に依るから，片方が単純ならば，もう片方がとても複雑であっても，$\sqrt{n}(\hat{\chi} - \chi(\eta))$ の漸近正規性が成り立つ．

例 1.4.10 (Rubin 2008)**.** 統計量 S_a, S_b を用いて，

$$\chi^1_\eta = S_{ab}a(Z)b(Z) + S_a a(Z) + S_b b(Z) + S_0 - \chi(\eta)$$

という形の影響関数を持つパラメータ $\chi(\eta)$ は mixed bias property を持つ．

例 1.4.11 (Chernozhukov 2018)**.** $a(Z) = E_\eta[Y|Z]$ と，$L_2(P_{\eta,Z}) \ni h \mapsto E_\eta[d(O,h)] \in \mathbb{R}$ が連続で affine 線型であるような d について，$\chi(\eta) = E_\eta[d(O,a)]$ と表せるパラメータ $\chi(\eta)$ は mixed bias property を持つ．

1.4.2　mixed bias property を持った影響関数の特徴付け

1.4.3　局外関数の特徴付け

1.4.4　例

1.4.5　Final remarks

1.5　De-Biased Machine Learning of Global and Local Parameters Using Regularized Riesz Representers

1.6　Double Debiased Machine Learning for Treatment and Structural Parameters

1.7　ON GAUSSIAN APPROXIMATION FOR M-ESTIMATOR

1.7.1　Introduction

(1) M 推定量 $\widehat{\theta}$ に分布の形が十分近いような，Gauss 過程の M-推定量 $\widehat{\theta}_G$ の構成法を提示する．
(2) 十分条件と，収束レートを考える．
(3) 乗数ブートストラップ法を与える．
(4) セミパラメトリックモデルと，non-Donsker なモデルとで同様のことを考える．
(5) 推定関数 ψ が滑らかでない場合も使える．

1.7.2　Gaussian Approximation for M-estimator

記法 1.7.1. $\epsilon \in (0,1]$ に対して，

(1) $H(\epsilon) := \log \mathcal{N}(\epsilon, \Theta, \|-\|)$ を計量エントロピーとする．
(2) $J(\epsilon) := \int_0^{\epsilon} \sqrt{1+H(\delta)}d\delta$ を計量エントロピー積分とする．

定理 1.7.2. 以上の仮定が成り立つとき，任意の可測集合 $A \subset \Theta, \epsilon \in (0,1]$ に対して，N が存在して，任意の $n \geqslant N$ に対して，

$$\exists_{C>0} \quad \left|\mathbb{P}_Z\left[\widehat{\theta} \in A\right] - \mathbb{P}_G\left[\widehat{\theta} \in A\right]\right| \leqslant C\left(\epsilon + \frac{H(\epsilon)}{n^{5/8}} + \frac{\Delta(n,\epsilon)}{\epsilon^{\kappa}}(\sqrt{H(\epsilon)}+1)\right).$$

1.8　GAUSSIAN APPROXIMATION OF SUPREMA OF EMPIRICAL PROCESSES

1.8.1　Introduction

記法 1.8.1. $X_1, X_2, \cdots$ が，確率変数 $(\Omega, \mathcal{F}, \mathbb{P}) \to (S, \mathcal{S})$ であって，$\mathbb{P}^{X_1} =: P$ に従うとする．$\mathcal{F}_n \subset \mathcal{L}(S)$ で添字付けられた経験過程

$$\left(\mathbb{G}_n f := \sqrt{n}(\mathbb{P}_n(f - Pf))\right)_{f\in\mathcal{F}_n}$$

の $l^{\infty}(\mathcal{L}(S))$ 上での収束を見る．この上限 $Z_n : S^n \to \mathbb{R}$ の過程 $(Z_n := \sup_{f\in\mathcal{F}_n} \mathbb{G}_n f)_{n\in\mathbb{N}}$ と法則同等になる中心化された Gauss 過程 $B_n : \Omega' \to l^{\infty}(\mathcal{L}(S))$ s.t. $Z_n \overset{d}{=} \sup_{f\in\mathcal{F}_n} B_n$ を満たすものを探すことを考える．このとき，各 B_n の共分散は

$$\mathrm{Cov}[B_n(f), B_n(g)] = E[B_n(f)B_n(g)] = \mathrm{Cov}[f(X_1), g(X_1)]$$

を満たすとする．

1.8.2　Abstract approximation theorem

定理 1.8.2. 次を仮定する.

(A1) $\mathcal{F} \subset l^\infty(\mathcal{L}(S))$ は各点可測である：ある可算部分集合 $\mathcal{G} \subset \mathcal{F}$ が存在して，$\forall_{f\in\mathcal{F}}\ \exists_{\{g_m\}\subset\mathcal{G}}\ \forall_{x\in S}\ g_m(x) \to f(x)$.
(A2) 包絡関数は積律有限：$\exists_{q\geqslant 3}\ F \in \mathcal{L}^q(P)$.
(A3) $\mathcal{F}$ は P-前 Gauss である：ある一様に緊密な中心化 Gauss 過程 $G_P : \Omega' \to l^\infty(\mathcal{F})$ が存在して，共分散が

$$\forall_{f,g\in\mathcal{F}} \quad E[G_P(f)G_P(g)] = P(fg) = E[f(X_1)g(X_1)]$$

を満たす.

このとき，$\kappa^3 \geqslant E[\|\mathbb{E}_n[|f(X_i)|^3]\|_{\mathcal{F}}]$ を満たすある正定数 $\kappa > 0$ が存在して，任意の $\epsilon \in (0,1]$ と $\gamma \in (0,1)$ に対して，$\widetilde{Z} \overset{d}{=} \sup_{f\in\mathcal{F}} G_P f$ が存在して，

$$\mathbb{P}\left[|Z - \widetilde{Z}| > K(q)\Delta_n(\epsilon,\gamma)\right] \leqslant \gamma\,(1 + \delta_n(\epsilon,\gamma)) + \frac{C\log n}{n}.$$

第 2 章

二重に頑健な推定量

(Chernozhukov 2016,[17]) では二重に頑健な推定量が構成できる推定関数のクラスが今までで一番一般的な枠組みで探求されている・(Rotnitzky and Robins et. al. 2020,[20]) では生物統計学者が，今までにない関数クラスを発見した．その第 1 段階について精査されているのが (Chernozhukov 2018[18], 2022[19]) である．(Seaman et. al., 2018 [21]) と (Bickel and Robins, 2001 [8]) が良い入門記事である．

2.1 定義

定義 2.1.1 (identifying estimating function / moment function, locally robust / orthogonal moment function, debiased GMM estimator, plug-in GMM estimator)**.** セミパラメトリックモデル $\mathcal{P} = (P_{(\theta,\gamma)})_{(\theta,\gamma)\in\Theta\times H}$ $(\Theta \subset \mathbb{R}^p, H \subset \mathrm{Map}(\Xi;\mathbb{R}))$ のパラメータ θ の推定を考える．

(1) 真値 $\theta_0 \in \Theta, \gamma_0 \in H$ に対して，$E[g(X,\gamma_0,\theta_0)] = 0$ かつ $\forall_{\theta\in\Theta\setminus\{\theta_0\}}\ E[g(X,\gamma_0,\theta)] \neq 0$ を満たす関数 $g : \mathcal{X} \times H \times \Theta \to \mathbb{R}^q$ を**推定関数** [8] 又は**識別可能積率関数** [17] という．
(2) このような推定関数が，**局所頑健**又は **(Neyman) 直交性を持つ** [17] とは，積率条件 g が第 1 段階推定量に対して微分が消えることをいう（第 1 段階のミスに対して頑健である）：
(3) 直交な積率関数から，cross-fitting を用いて構成された標本積率方程式の解として得られる GMM 推定量を，**脱バイアス化された推定量**という．一方で，通常の識別可能な積率関数を用いて構成された推定量を**代入推定量**という．

要諦 2.1.2. 通常，モーメント関数 g は，第 1 段階で推定される未知関数 γ を含む．そこで推定量 $\hat{\gamma}$ を用いて

$$\frac{1}{n}\sum_{i=1}^{n} g(x_i,\hat{\gamma},\theta) = 0$$

を解くことになるが，$\hat{\gamma}$ の一致推定に失敗すると，θ は決して見つからないことになる．そこでモーメント関数 g に対して，第 1 段階のミスに関する影響関数を加えることで，直交性を持つモーメント関数が構成できる，という理論である．第 1 段階のミスはパラメトリックに起きるわけではないので，必然的にセミパラメトリックな理論となる．

利点 2.1.3 (plug-in GMM と比較した際の美点)**.**

(1) 第 1 段階でモデル選択がなされた場合でも，debiased GMM の信頼区間は有効なままである．これは機械学習を第 1 段階に採用した際にも当てはまる．
(2) 正則化された第 1 段階を採用する 2 段階推定において，debiased GMM 推定量は $\sqrt{n}$-一致性を持つ．
(3) 直交積率関数が第 1 段階について affine ならば，debiased GMM 推定量は二重に頑健になる．
(4) debiased GMM は，剰余項がより速く収束し，平均二乗誤差も 2 次の微小項について減少している．
(5) debiased GMM を構成するための正則性条件は，通常の積率関数に課される条件よりも一般的で，単純明快である．
(6) debiased GMM は，第 1 段階推定量の L^2-一致性とその収束の速さに関する条件の下で，漸近正規性を持つ．

2.2　議論

記法 2.2.1.

(1) 真の分布が F であるとき，標本数 n を大きくすると，$\widehat{\gamma} \xrightarrow{n\to\infty} \gamma(F)$ なる確率収束極限を持つとする（正則性条件その 1 ということで R1 と呼ぶ）．$\gamma(F) = \gamma_0$ のとき，$\widehat{\gamma}$ は弱一致推定量である．

(2) $\mu(F) := E[g(X, \mu(F), \theta)]$ なる統計的汎関数の (ノンパラメトリック) 影響関数の定義には，Gateaux 微分としての特徴づけ

$$\frac{d}{dt} E[g(X, \gamma(F_t), \theta)]\Big|_{t=0} = \int_{\mathfrak{X}} \phi(x, \gamma_0, \alpha_0, \theta) H(dx), \quad E[\phi(X, \gamma_0, \alpha_0, \theta)] = 0.$$

ただし，F_t は真の分布 F_0 を H 方向に接道したもので，α は追加で作った自由度である．ϕ は第 1 段階の γ の推定がミスったときの，第 2 段階の積率方程式の漸近的な振る舞いに与える影響（微分係数）を表している．またおまけで，α についての影響も取り入れている．

定義 2.2.2 (orthogonal moment function, debiased sample moment function)**.**

(1) $\psi(X, \gamma, \alpha, \theta) := g(X, \gamma, \theta) + \phi(X, \gamma, \alpha, \theta)$ を**直交モーメント関数**という．

(2) **脱バイアス化された標本モーメント関数**とは，$\widehat{\psi}(\theta) := \widehat{g}(\theta) + \widehat{\phi}$ をいう．ただし，

$$\widehat{g}(\theta) := \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in I_k} g(X_i, \widehat{\gamma}_i, \theta), \quad \widehat{\phi} := \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in I_k} \phi(X_i, \widehat{\gamma}_i, \widehat{\alpha}_i, \widehat{\theta}_i).$$

(3) $\widehat{\theta} := \arg\min_{\theta \in \Theta} \widehat{\psi}(\theta)^\top \widehat{\Psi}^{-1} \widehat{\psi}(\theta)$ は脱バイアス化された GMM 推定量となる．

2.3　背景

計量経済学，生物統計学などのあらゆる分野において，因果推論の文脈＝欠測データの文脈で，ノンパラメトリック又は高次元な第 1 段階を含んだ 2 段階推定モデルが採用されているのが現状である (Marschak's Maxim の影響だろうか？)．このときの第 1 段階の推定について，2 つのモデルを立てて，どちらかで正解すれば，この 2 段階推定法で得られる推定量が一致性を持つ，という性質を考える．また，この直交性を用いて，ちょっとした適合度検定を行える．

2.3.1　機械学習について

共変量や状態変数が高次元であるとき，機械学習の採用が望まれる．Lasso, Dantzig, neural nets, boosting など．正則化とは，特に Bayes の観点からはモデルについての事前情報を追加することをいう．これはモデル選択同様，第 1 段階にバイアスを増やすことになる．そこで，Debiased で抜かなきゃいけないことになる．

2.3.2　モーメント法について

Karl Pearson 1894 による．

2.3.3　一般化モーメント法について

Generalized Method of Moments

Lars Hansen 1984 による．この手法の開発とファイナンスへの応用より，2013 年にノーベル経済学賞が授与された．このクラスの推定量は，一致性と漸近正規性を持ち，さらにモーメント条件以外の情報を使わずに構成される全ての推定量の中で漸近最適である．

要諦 2.3.1. 母集団分布の積率について，一種の「直交条件」を仮定する．標本分布に関してその条件を当てはめることで推定量を構成する手法を GMM という [32].

記法 2.3.2. 過程 $(x_t)_{t\in\mathbb{N}}$ の実現を観測するとする．このような離散過程全体の空間 $\mathcal{L}(\mathbb{N})$ 上の確率分布の族を $\mathcal{M}$ とする．$\Theta\subset\mathbb{R}^k$ を局外母数の空間とする．この過程は，$f:\mathcal{L}(\mathbb{N})\times\Theta\to\mathbb{R}^r$ を用いて次のように表された積率条件：

$$\exists_{\beta_0\in\Theta}\ \forall_{x_t\in\mathcal{L}(\mathbb{N})}\quad E[f(x_t,\beta_0)]=0.$$

と識別可能性条件 $E[f(x_t,\beta)]=0\Leftrightarrow\beta\neq\beta_0\in\Theta$ を満たすとする．このとき，目標のパラメータ β は有限次元であるが，そのほかに分布について仮定を置いていないので，モデルはセミパラメトリックになっている．

観測 $(x_t)_{t\in\mathbb{N}}$ のそれぞれについて，関数 $g:\Theta\to\mathbb{R}^r$ が，$\mathbb{P}_n$ が $\mathbb{R}$ 上のサイズ N の経験分布として，$g(\beta):=E_{\mathbb{P}_n}[f(x_t,\beta)]$ すなわち

$$g(\beta):=\frac{1}{N}\sum_{n=1}^{N}f(x_t,\beta)$$

定義 2.3.3. 次のような方法で構成される β の推定量 $b_N:\mathbb{R}^N\to\Theta$ を**一般化モーメント推定量**という．

(1) 正定値な荷重行列 $W\in M_r(\mathbb{R})$ に対して，$b_N:=\arg\min_{\beta\in\Theta}g_N(\beta)^\top Wg_N(\beta)$．$N\to\infty$ のとき，大数の法則より $g_N\to E[f]$ であり，このとき $b_\infty=\beta_0$ を満たすことが根拠となっている．実際，一致性と漸近正規性を持つ．
(2) $A\in M_r(\mathbb{R})$ を選択行列 (selection matrix) とし，$Ag_N(\beta)=0$ の解とする．

2.3.4　不完全データ＝因果推論の面

欠損データの扱い

(1) 逆傾向スコア荷重 (IPW strategy)：欠損データの従う確率分布を推定する．
(2) 代入法 (inputation strategy)：欠損データそのものをモデル (imputation model) する．特に主流な多重代入法 (multiple inputation) は最尤法に最も近い．

(2) の方が大抵の場合効率的であるが，model misspecification を検出するすべがない．この 2 つを結合して，2 段階構造にしたのが二重にロバストな手法である．AIPW とは，imputation model を導入して，misspecification がない場合に IPW よりも効率を上げた最適なものである．1999 年に Scharfstein が，missingness model が正しければ一致性を持つだけでなく，たとえ missingness model がミスっても imputation model が正しければやはり一致性をもつ性質を発見した．(1) より効率的で，(2) よりも model misspecification に関して頑健である．

注 2.3.4 (代入法とは？)**.** 欠測データに関する対処として list-wise 削除がある．使用するデータを取捨選択しているわけなので，残ったデータにはバイアスが残るが，最も簡単なので利用されてしまう．そこで，削除するのではなく，穴を埋めることを考える．

(1) ホットデッキ代入法：無作為に抽出された類似の記録を打ち込む．ホットとは処理中のパンチカードの山のことで，ここから再利用することを指す．LOCF(last observation carried forward) は，最後の観測値の繰り返しである．
(2) コンピュータの性能が向上すると，過去の類似した調査から考えて埋め込む．これをコールドデッキ代入という．
(3) 平均値置換．
(4) 非負行列因子分解：天文学分野で画像データの欠損処理．
(5) 回帰代入法：他の変数から欠損データを推測する回帰モデルを立てる．
(6) 多重代入法：Rubin による．
   i. 無作為にデータから 1 つ選んで欠損データに代入する（単一代入法）．これを m 回繰り返してコピーを作る．
   ii. m 個をそれぞれ分析する．
   iii. pooling：関心のある変数の平均・分散・信頼区間を計算し，m 個の結果を 1 つにまとめる．
   この手法は MNAR でも使える．マルコフ連鎖モンテカルロ法が最もよく用いられる (MICE: multiple imputation by chained equations).

2.3.5　IPW, RI, AIPW

記法 2.3.5. サイズ n のデータ Y は欠損があり，$\beta := E[Y]$ の推定を考える．二次的変数 W には欠損がないとする．より一般の場合は次の節で考える．Y_i, W_i で i-単位を表す．$R : [n] \to 2$ を，Y_i が観測されたかの真理値を表す．$R_i = 1$ である個人を完全な事例 (complete case) という．$(W_1, Y_1, R_1), \cdots, (W_n, Y_n, R_n)$ は i.i.d. とするので，添字は省略可能．

議論 2.3.6. 当然，欠損がない場合は $n^{-1} \sum_{i=1}^{n} Y_i$ が推定量である．では単純に修正した

$$\frac{\sum_{i=1}^{n} R_i Y_i}{\sum^{n} i = 1 R_i}$$

は，R が Y と独立でない限り (MCAR) 一致性を持たない．現実のデータは無作為割当ばかりではないので，この条件を弱めて，MAR，すなわち，$Y \perp\!\!\!\perp R | W$ を仮定した下で一致性を持つ推定量 IPW, RI, DR を考える．

2.3.5.1　IPW

$\pi(W) := P[R = 1|W]$ として，パラメトリックな部分を抽出することは自然である．W の値を持った個人を $\pi(W)^{-1}$ によって荷重する．欠測する可能性の高い W を持った個人ほど，$\pi(W)^{-1}$ 人分として多く数える測度変換を行う．すると，この Y のうちだけ complete case を抽出して，$\pi(W)^{-1}$ によって荷重して得られる $Y\pi(W)^{-1}$ の分布は，W の値も，Y の値も，欠損前の母集団の分布と同じになる．よって，先ほどの推定量を

$$n^{-1} \sum_{i=1}^{n} R_i \pi(W_i)^{-1} Y_i, \quad \frac{\sum_{i=1}^{n} R_i \pi(W_i)^{-1} Y_i}{\sum_{i=1}^{n} R_i \pi(W_i)^{-1}}$$

で置き換える．前者を IPW，後者を IPW,B などといい，後者の値は標本の Y の値域に必ず入る (sample-bounded)．実験計画がなされた場合を除いて，$\pi(W)$ は推定する必要がある．これはパラメトリックに $\pi(W; \alpha)$ によっておこない，これを**欠損モデル (missingness model)** という．まずは α をデータ $(W_1, Y_1, R_1), \cdots, (W_n, Y_n, R_n)$ から推定することになる．機械学習によっても良い．

定理 2.3.7. 次が成り立つ場合，$\widehat{\beta}_{IPW}, \widehat{\beta}_{OPW,B}$ は β の一致推定量である．

(1) モデル $\pi(W; \alpha)$ は正しい (correctly specified)
(2) $\widehat{\alpha}$ は α の一致推定量である．
(3) positivity : $\exists_{\delta > 0}\ P[\pi(W) \geqslant \delta] = 1$．すなわち，ある W であって，これに属する個人の Y がすべて欠測していることはないものとする．
(4) $\alpha \mapsto \pi(W; \alpha)$ は十分に滑らかである．

注 2.3.8. 一部のデータを引き伸ばして使っているので，分散は大きくなる．が，R, W という完全なデータのみを使っているので，適合度検定は簡単にでき，モデルの誤設定はより簡単に診断できる．

2.3.5.2　RI: Regression Imputation

これはセミパラメトリック手法特有のものか！？なぜ議論が有限に落ちているのか．すごいな．

$E[Y|W]$ の分布をモデル $m(W; \gamma)$ によって考える．これを結果モデル (outcome model) という．完全なケースから推定量 $\widehat{\gamma}$ を割り出し，

$$\widehat{\beta}_{RI}(\widehat{\gamma}) := n^{-1} \sum_{i=1}^{n} m(W_i; \widehat{\gamma})$$

とする．

定理 2.3.9. 次が，$\widehat{\beta}_{RI}$ が一致性を持つ十分条件である．

(1) 結果モデル $m(W; \gamma)$ が正しく設定されている．

(2) $\hat{\gamma}$ は一致性を持つ.

また，$\hat{\gamma}$ が最適であるならば，$\hat{\beta}_{RI}$ も最適である.

注 2.3.10 (本当に回帰代入法になっているのか？). モデル $m(W;\gamma)$ が切片 (intercept) を持つ標準的な一般化線形モデルで，$\hat{\gamma}$ が機械学習推定値であるとき，

$$\sum_{i=1}^{n} R_i m(W_i;\hat{\gamma}) = \sum_{i=1}^{n} R_i Y_i$$

で，

$$\hat{\beta}_{RI} = n^{-1}\sum_{i=1}^{n}\{R_i Y_i + (1-R_i)m(W_i;\hat{\gamma})\}$$

と表せる．これは，欠測値に $m(W;\hat{\gamma})$ を代入して平均を取っているということになる.

注 2.3.11. 最適性は，(1) のモデル設定の難関を通過して初めて得られる．特に，欠測している場合といない場合で共変量 W の分布が大きく変わる場合は，欠測していないデータを使っている回帰なので，これを用いて外挿することは極めてリスキーである．特に，モデルの誤設定の診断さえ難しい.

2.3.5.3　AIPW

欠測データをより活用することで，IPW の分散を減らすことが出来る.

$$\hat{\beta}_{DR}(\hat{\alpha},\hat{\gamma}) = \frac{1}{n}\sum_{i=1}^{n}\frac{R_i}{\pi(W_i;\hat{\alpha})}Y_i + \frac{1}{n}\sum_{i=1}^{n}\left(1-\frac{R}{\pi(W_i;\hat{\alpha})}\right)m(W_i;\hat{\gamma}).$$

第一項は IPW に他ならず，第二項を augmented term という．また，

$$\hat{\beta}_{DR}(\hat{\alpha},\hat{\gamma}) = \frac{1}{n}\sum_{i=1}^{n} m(W_i;\hat{\gamma}) + \frac{1}{n}\sum_{i=1}^{n}\frac{R_i}{\pi(W_i;\hat{\alpha})}\left(Y_i - m(W_i;\hat{\gamma})\right)$$

と表示すると，第一項は RI に他ならず，第二項で補正していると見れる.

定理 2.3.12 (2 重ロバスト性). 次の条件のいずれかが満たされたとき，$\hat{\beta}_{DR}$ は一致性と漸近正規性を持つ.

(1) 欠測モデル $\pi(W;\alpha)$ が正しく設定され，$\hat{\alpha}$ は一致推定量である.
(2) 結果モデル $m(W;\gamma)$ が正しく設定され，$\hat{\gamma}$ は一致推定量である.

［証明］.

(1) augmented term が 0 に収束するため.
(2) 補正項が 0 に収束するため.

■

2.3.6　セミパラメトリック理論

「パラメトリック推定を 2 回行えば良い」という枠組みはセミパラメトリックになっている.

2.4　INFERENCE FOR SEMIPARAMETRIC MODELS

Bickel, P. J., and Jaimyoung Kwon. (2001). INFERENCE FOR SEMIPARAMETRIC MODELS: SOME QUESTIONS AND AN ANSWER.

2.4.1 Introduction

歴史を説明している．1975s から non-/semi-parametric model が反映した理由は，コンピュータの台頭と，

(1) ノンパラの翻訳不可能性の解消．

例 **2.4.1.** 計量経済学における指数モデルなどは，セミパラメトリック回帰の例に入る．

第 3 章

参考文献

参考文献

[1] 吉田朋広. (2006). 数理統計学. 朝倉書店.

[2] 芝村良. (2004).『R. A. フィッシャーの統計理論』. 九州大学出版会.

[3] van der Vaart, A. W. (1998) *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge.

[4] Bolthausen, E., Vaart, A., and Perkins, E. (2002). *Lectures on Probability Theory and Statistics*. Springer Berlin, Heidelberg.

[5] Kennedy, E. H. (2022). Semiparametric Doubly Robust Targeted Double Machine Learning: A Review. arXiv: 2203.06469.

[6] Kennedy, E. H., Balakrishnan, S., and G' Sell, M. (2020). Sharp instruments for classifying compliers and generalizing causal effects. *The Annals of Statistics*. 48(4): 2008 - 2030.

[7] Bickel, P. J., Chris A.J. Klaassen, Ya'acov Ritov, Wellner, J. A. (1998) *Efficient and Adaptive Estimation for Semiparametric Models*. Springer New York.

[8] Bickel, P. J., Kwon, J. (2001). Inference for Semiparametric Models: Some Questions and an Answer. *Statistica Sinica*. 11: 863-960.

[9] Hampel, F. R. (1968). *Contributions to the theory of robust estimation*. Unpublished dissertation, Berkeley: University of California.

[10] Hampel, F. R. (1971). A General Qualitative Definition of Robustness. *The Annals of Mathematical Statistics*. 42(6): 1887-1896.

[11] Hampel, F. R. (1974). The Influence Curve and its Role in Robust Estimation. *Journal of the American Statistical Association*, 69(346): 383-393. DOI: 10.1080/01621459.1974.10482962.

[12] Bliss, G. A. (1915). A Note on Functions of Lines. *Proceedings of the National Academy of Sciences of the United States of America*. 1(3): 173-177.

[13] Vito Volterra. (1913). *Leçons sur les fonctions de lignes*. Paris: Gauthier-Villars.

[14] von Mises, R. (1947). On the Asymptotic Distribution of Differentiable Statistical Functions. *The Annals of Mathematical Statistics*. 18(3): 309-348. DOI: 10.1214/aoms/1177730385.

[15] Newey, W. K., and McFadden, D. (1994). Large Sample Estimation and Hypothesis Testing. *Handbook of Econometrics*. Volume 4. North-Holland, an imprint of Elsevier, Amsterdam.

[16] Ichimura, H., and Newey, W. K. (2022). The Influence Function of Semiparametric Estimators. *Quantitative Economics*. 13(1): 29-61. DOI: 10.3982/QE826.

[17] Chernozhukov, V., Escanciano, J.C., Ichimura, H., Newey, W. K., and Robins, J. M. (2016). Locally Robust Semiparametric Estimation. arXiv:1608.00033.

[18] Chernozhukov, J., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. K., and Robins, M. J. (2018). Double/Debiased Machine Learning for Treatment and Structural Parameters. *The Econometrics Journal*. 21(1): C1-C68.

[19] Chernozhukov, V., Newey, W. K., and Singh, R. (2022). DeBiased Machine Learning of Global and Local Parameters Using Regularized Riesz Representers. *The Econometrics Journal*. utac002, 00: 1-26.

[20] Rotnitzky, A., Smucler, E., and Robins, J. M. (2020). Characterization of parameters with a mixed bias property. *Biometrika*. 108(1): 231-238.

[21] Seaman, S. R., and Vansteelandt, S. (2018). Introduction to Double Robust Methods for Incomplete Data. *Statistical*

Science. 33(2): 184-197.

[22] Hines, O., Dukes, O., Diaz-Ordaz, K., and Vansteelandt, S. (2022). Demystifying statistical learning based on efficient influence functions. *The American Statistician*. DOI: 10.1080/00031305.2021.2021984.

[23] Malliavin, P., and Cruzeiro, A. B. (1996). Renormalized Differential Geometry on Path Space: Structural Equation, Curvature. *Journal of Functional Analysis*. 139: 119-181.

[24] Shinzo Watanabe. (1987). Analysis of Wiener Functionals (Malliavin Calculus) and its Applications to Heat Kernels. *Annals of Probability*. 15(1): 1-39. DOI: 10.1214/aop/1176992255.

[25] Pedersen, G. K. (1989). *Analysis Now*. Springer New York.

[26] van der Laan, M. J., Rose, S. (2011). *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer New York.

[27] Robins, J. M., and Ritov, Y. (1997). Toward a Curse of Dimensionality Appropriate (CODA) Asymptotic Theory For Semi-parametric Models. *Statistics in Medicine*. 16: 285-319.

[28] Robins, J. M., Li, L., Tchetgen, E., and van der Vaart, A. W. (2008). Higher order influence functions and minimax estimation of nonlinear functionals. *IMS Collections: Probability and Statistics: Essays in Honor of David A. Freedman*. 2: 335 - 421.

[29] van der Vaart, A. W. (2014). Higher Order Tangent Spaces and Influence Functions. *Statistical Science*. 29(4), Special Issue on Semiparametrics and Causal Inference: 679-686.

[30] Imaizumi, M., Otsu, T. (2020). On Gaussian Approximation for M-estimator. arXiv:2012.15678.

[31] Chernozhukov, V., Chetverikov, D., Kato, K. (2014). Gaussian approximation of suprema of empirical processes. *The Annals of Statistics*. 42(4): 1564-1597.

[32] Hansen, L. P. (2007). Generalized Method of Moments Estimation. http://home.uchicago.edu/~lhansen/palgrave.pdf