高次元統計学におけるリサンプリング法に対する統計力学的アプローチ

東京大学大学院 理学系研究科付属 知の物理学研究センター
髙橋 昂 (Takashi TAKAHASHI)

2020 年 10 月 29 日 (木)13:00-14:00 @ Zoom meeting

» 今日の話の概略

とある統計手法の性質の、統計力学風の解析

» 今日の話の概略

とある統計手法の性質の、統計力学風の解析

* 高次元統計学での変数選択
  - サンプル数/パラメータ数 $\to$ 有限
  - 統計モデルで relevant な変数はどれか
  - バイオインフォマティクスなどで現れる

» 今日の話の概略

とある統計手法の性質の、統計力学風の解析

* 高次元統計学での変数選択
  - サンプル数/パラメータ数 $\to$ 有限
  - 統計モデルで relevant な変数はどれか
  - バイオインフォマティクスなどで現れる
* リサンプリング法を用いた変数選択法
  - LASSO 等のスパース推定法の改良
  - 取得データに関する統計性 $+\alpha$ を利用
  - 性質若干非自明

» 今日の話の概略

とある統計手法の性質の、統計力学風の解析

* 高次元統計学での変数選択
  - サンプル数/パラメータ数 $\to$ 有限
  - 統計モデルで relevant な変数はどれか
  - バイオインフォマティクスなどで現れる
* リサンプリング法を用いた変数選択法
  - LASSO 等のスパース推定法の改良
  - 取得データに関する統計性 $+\alpha$ を利用
  - 性質若干非自明
* その統計力学的な解析
  - ランダム問題
  - 平均場漸近論（サンプル数, パラメータ数 $\to \infty$, サンプル数/パラメータ数 $\to (0, \infty)$）
  - レプリカ法, ランダム行列のスペクトルの関数...

» 本題に入る前に、自己紹介: 略歴

* 2013: 東京大学 教養学部 基礎科学科卒業
* 2013-2015: 東京大学大学院 総合文化研究科広域科学専攻 修士課程
  - 指導教員: 福島孝治
  - テーマ：モンテカルロ・シミュレーションによる３次元ポッツグラスの研究
    - 物性物理の、統計力学的側面の研究
* 2015-2017: 株式会社構造計画研究所 創造工学部
  - マルチエージェントシミュレーションとか統計分析 (のお手伝い)
* 2017-2020: 東京工業大学情報理工学院 数理・計算科学専攻 博士課程
  - 指導教員：樺島祥介
  - テーマ：高次元統計学に対する統計力学的アプローチ
    - 統計力学で、統計学の一部の研究
* 2020-: 東京大学 理学系研究科付属知の物理学研究センター 樺島研 PD

» 高次元線形モデルの変数選択

* データ
  - $D = \{(\boldsymbol{a}_\mu, y_\mu)\}_{\mu=1}^M$,
  - $\boldsymbol{a}_\mu \in \mathbb{R}^N$: 特徴量／予測子, $y_\mu \in \mathbb{R}$: 応答変数
* 観測プロセス
  - $y_\mu = \boldsymbol{a}_\mu^\top \boldsymbol{x}_0 + \epsilon_\mu, \mu = 1, 2, \ldots, M$
  - $\boldsymbol{x}_0 \in \mathbb{R}^N$: 真のパラメータ
    - パラメータのスパース性は仮定 $\dfrac{\#\{i \mid x_{0,i} \neq 0\}}{N} = \rho \in (0, 1)$
  - $\epsilon_\mu \sim \mathcal{N}(0, \sigma^2)$: 観測ノイズ
* 高次元性
  - $\alpha \equiv M/N \in (0, \infty)$, 特に、$\alpha < 1 \Leftrightarrow$ 古典的な統計学 $M \gg N$
* 変数選択のゴール

サポート $\mathrm{supp}(\boldsymbol{x}_0) \equiv \{i \mid x_{0,i} \neq 0\}$ をデータ D から推定すること

$\Leftrightarrow$ 2 乗誤差の良さ $\|\hat{\boldsymbol{x}} - \boldsymbol{x}_0\|_2^2/N$ （for some estimator $\hat{\boldsymbol{x}}$）

» 例: リボフラビン生成率の予測 [BKM14]

* $\boldsymbol{a}_\mu \in \mathbb{R}^N$: 人間の遺伝子発現データ
* $y_\mu \in \mathbb{R}$: リボフラビン (ビタミン B_2 らしい) の生成レート

* ゴール：どの遺伝子が大事かを知りたい
* サンプル数 $M = 71$,
* 特徴量／予測子の次元 $N = 4088$

* $\boldsymbol{a}$ という入力点に y という出力があるイメージ...
* $\boldsymbol{a}$ は計画的に規則正しく得られている場合もあれば、ランダムにサンプリングしている場合もある

» 素朴なスパース推定

* LASSO[Tib96]

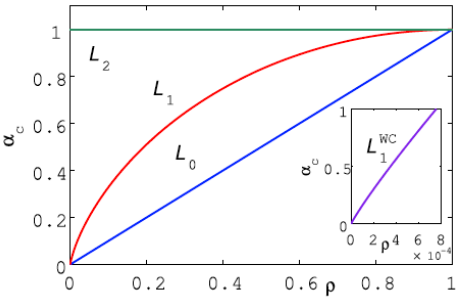
$$\hat{\boldsymbol{x}}(D,\lambda) = \arg\min_{\boldsymbol{x}\in\mathbb{R}^N} \left[\frac{1}{2}\sum_{\mu=1}^{M}(y_\mu - \boldsymbol{a}_\mu^\top \boldsymbol{x})^2 + \sum_{i=1}^{N}\lambda|x_i| \right]$$

* 凸最適化
  - 推定は計算量的に容易
* 推定量はスパース
  - 十分大きな λ に対しては $\hat{\boldsymbol{x}}(D,\lambda)$ の多くの成分はゼロ
  - 回帰 + 変数選択 の同時実行

» LASSO 推定量の良さ

* 平均場漸近論
  $M, N \to \infty, \alpha = M/N \in (0, \infty)$ での相転移 [KWT09, DMM09]
  - $\sigma = 0$ の場合、相境界線より上で

$$\|\hat{\boldsymbol{x}} - \boldsymbol{x}_0\|_2^2 \to 0$$



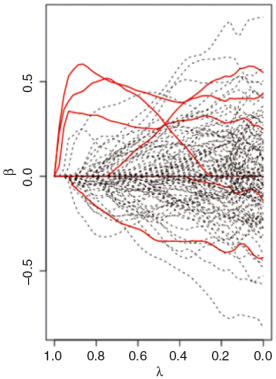
* より一般的なバウンド（適当な λ と A の仮定のもとで）[BVDG11]

$$s_0 \equiv \rho N, M \geq s_0 \log N \Rightarrow \|\hat{\boldsymbol{x}} - \boldsymbol{x}_0\|_2^2 \leq c \frac{s_0}{M} \log N,$$

  - s_0 が小さければ普通の線形回帰のレートに $\log N$ の補正.

» LASSO 推定量の悪さ

* 点推定. false-positive（$\hat{x}_i \neq 0$ and $x_{0,i} = 0$）がコントロールできない[1]
* 正則化パラメータ λ の選択の問題
  - 実データでは LASSO 解のパス $\{\hat{\boldsymbol{x}}(D, \lambda) \mid \lambda \in \Lambda\}$ はノイジー
  - 変数選択に適切な λ？
    - $\Leftrightarrow$ 予測誤差の問題（交差検証で OK）



* riboflavin data での半合成実験 [MB10]
  - 赤: true positive
  - 黒: false positive

[1] 最近 post-selection 推論が整備されてきているのでそうでもない気がするが

» リサンプリングを用いた安定性選択法

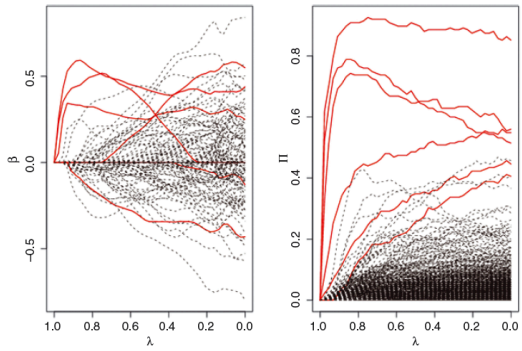
* ブートストラップリサンプリングデータ
  - $D^* \equiv \{(\boldsymbol{a}^*_\mu, y^*_\mu)\}_{\mu=1}^{M_\mathrm{B}}$
    - $D = \{(\boldsymbol{a}_\mu, y_\mu)\}_{\mu=1}^{M}$ からの重複ありで M_B 点抽出したデータ
    - 経験分布を利用したシミュレーション
* randomized LASSO
  - $\lambda_i \sim (1-p_w)\delta(\lambda-\lambda_0) + p_w\delta(\lambda - w\lambda_0)$
* 安定性 $\boldsymbol{\Pi}$
  - $\Pi_i(D,\lambda_0) \equiv \mathrm{Prob}_{D^*,\lambda}[\hat{x}_i(D^*,\boldsymbol{\lambda}) \neq 0]$,
  - $\hat{\boldsymbol{x}}(D^*,\boldsymbol{\lambda}) = \arg\min_{\boldsymbol{x}\in\mathbb{R}^N}\left[\frac{1}{2}\sum_{\mu=1}^{M_\mathrm{B}}(y^*_\mu - (\boldsymbol{a}^*_\mu)^\top\boldsymbol{x}) + \sum_{i=1}^{N}\lambda_i|x_i|\right]$

  ※ 実際的にはモンテカルロ近似する

  $\Pi_i \simeq \frac{1}{B}\sum_{b=1}^{B}\mathbb{1}(\hat{x}_i(D^*_b,\boldsymbol{\lambda}^*_b)\neq 0)$, for some value of B
* 安定性選択（stability selection: SS）法 [MB10] に基づくサポートの推定
  - $\{i \mid \max_{\lambda_0\in\Lambda}(\Pi_i(D,\lambda_0)) \geq \Pi_\mathrm{th}\}$ for some $\Pi_\mathrm{th} \in (0,1)$

» SS 法の性質

* 安定性 Π のパスはずっと安定的
  - 正則化パラメータの選択に less sensitive



* 左: LASSO 解 $\hat{\boldsymbol{x}}(D, \lambda)$ のパス
* 右: 安定性 $\Pi(D, \lambda_0)$ のパス

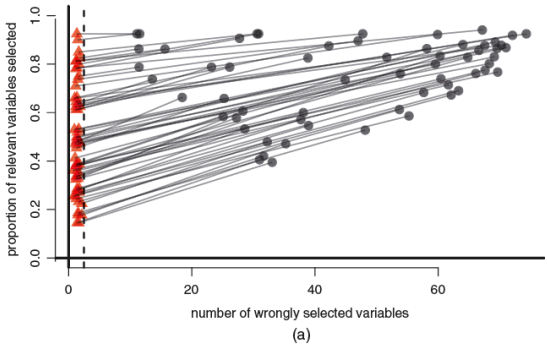
[MB10] から引用

» SS 法の性質 (cont.)

* $\Pi_{\text{th}} \in (1/2, 1)$ での偽陽性の個数 V に対するバウンド [MB10]

$$\mathbb{E}[V] \leq \frac{1}{2\Pi_{\text{th}} - 1} \frac{q_\Lambda^2}{N}$$

  - Λ: 解パスの λ の集合
    ※1 点でも OK (point-wise control)
  - q_Λ: SS 法で選択された変数の個数



* 点線: $\mathbb{E}[V] \leq 2.5$ のバウンド
* 黒丸: 交差検証で λ を選んだ場合
* 赤: SS で刈り込んだ後

» 興味

* 理論的側面 ← 今日の話
  - false positive を減らしていった時、true positive の発見はきちんと残るのか？残るのならそれはどういうときに…
  - リサンプリングサンプル数 M_{B} にはどう依存するんだろう…
  - 何か平均場解析してみたい
* 計算量的側面
  - 何度も推定を繰り返す必要がある
  - 最適化問題を繰り返し解くのはコストが大きい可能性があるのでサボれないか
* 参考文献
  - arXiv:2003.08670 [TK20b]
    - 近似アルゴリズムと、平均場漸近論の表式だけ
  - arXiv:2001.02824 [TK20a]
    - appendix にレプリカ計算の概略…

» ここまで整理

* 高次元線形モデルの変数選択
  - モデルパラメータの非ゼロ成分のサポートの推定
  - バイオインフォマティクスとかで現れる
  - データ数はパラメータ数と比較して多くない
* 素朴なスパース推定法
  - 簡便
  - 変数選択にはそれ単体ではあまりよくない
    - false positive のコントロールの問題
    - 正則化パラメータの選択の問題
* リサンプリング法との組み合わせ
  - 素朴なスパース推定法の問題点を改善
  - 他の問題も発生する
    - 理論的側面
    - 計算量的側面

イントロダクション

リサンプリング法の平均場漸近論

リサンプリング法の平均場での性質

まとめ

» 準備 1/2: 問題設定／ノーテーション整理

* 特徴量行列 $A = [a_{\mu i}]_{\substack{1\leq\mu\leq M\\ 1\leq i\leq N}} \in \mathbb{R}^{M\times N}$ が回転不変行列と仮定

  $A \overset{\text{SVD}}{=} USV^\top$ に対し、U, V が直交行列上の一様分布
* データ生成モデル
  - $y_\mu \sim \mathcal{N}(\boldsymbol{a}_\mu^\top \boldsymbol{x}_0, \sigma^2) \equiv q_{y|z}(y_\mu | \boldsymbol{a}_\mu^\top \boldsymbol{x}_0)$
  - $x_{0,i} \sim q_{x_0}(x_0) = \rho \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x_{0,i}^2} + (1-\rho)\delta(x_{0,i})$
* 平均場漸近論
  - $M, N \to \infty, M/N \to \alpha \in (0, \infty)$
* ノーテーション
  - M: サンプル数
  - N: パラメータ数
  - α: サンプル数とパラメータ数の比

» 準備 2/2: SS 法の統計力学的定式化

* Boltzmann 分布

$$p^{(\beta)}(\boldsymbol{x}, \boldsymbol{z}) = \frac{1}{Z} \prod_{\mu=1}^{M} p_{y|z}(y_\mu|z_\mu)^{\beta c_\mu} \prod_{i=1}^{N} e^{-\beta\lambda_i|x_i|} \delta(\boldsymbol{z} - A\boldsymbol{x})$$

  - $p_{y|z}(y_\mu|z_\mu) \propto \exp(-\frac{1}{2}(y_\mu - z_\mu)^2)$: 尤度みたいなもの
  - $\boldsymbol{z} \in \mathbb{R}^M$: 補助変数
  - c_μ: データ点 $(\boldsymbol{a}_\mu, y_\mu)$ が D^* に含まれていた個数
    - ブートストラップサンプル D^* は経験分布からのサンプルなので、$\boldsymbol{c}$ と情報等価
  - $M \gg 1$ では独立なポアソン分布で近似可: $c_\mu \sim e^{-1}/c_\mu!$
  - $\lambda_i \sim p_\lambda(\lambda) = (1 - p_w)\delta(\lambda - \lambda_0) + p_w\delta(\lambda - w\lambda_0)$

* 安定性
  - $\Pi_i = \mathbb{E}_{\boldsymbol{c},\boldsymbol{\lambda}}\left[\mathbb{1}\left(\hat{x}_i(\boldsymbol{c}, \boldsymbol{\lambda}) \neq 0\right)\right]$
  - $\hat{x}_i(\boldsymbol{c}, \boldsymbol{\lambda}) = \lim_{\beta\to\infty} \int \boldsymbol{x} p^{(\beta)}(\boldsymbol{x}, \boldsymbol{z}) d\boldsymbol{x} d\boldsymbol{z}$

» リサンプリング統計量の計算の難しさ

- 一般に $\hat{x}_i(\boldsymbol{c}, \boldsymbol{\lambda})$ の $r \in \mathbb{N}$ 次モーメントの計算を考えてみる
- 定義から

$$\mathbb{E}_{\boldsymbol{c},\boldsymbol{\lambda}}[\hat{x}_i^r] = \lim_{\beta\to\infty} \int \prod_{s=1}^{r} x_{s,i} \prod_{s=1}^{r} \mathbb{E}_{\boldsymbol{c},\boldsymbol{\lambda}} \left[\frac{1}{Z(\boldsymbol{c},\boldsymbol{\lambda})} \prod_{\mu=1}^{M} p_{y|z}(y_\mu|z_{s,\mu})^{\beta c_\mu} \prod_{i=1}^{N} e^{-\beta\lambda_i|x_{s,i}|} \delta(\boldsymbol{z}_s - A\boldsymbol{x}_s) \right] d^r\boldsymbol{x} d^r\boldsymbol{z}$$

- Z^{-r}の因子が取り扱い困難
  - これがわかるような問題なら困難はあまりなさそうだが、わからない

» レプリカ法を通じた問題の書き換え

$$\mathbb{E}_{\boldsymbol{c},\boldsymbol{\lambda}}[\hat{x}_i^r] = \lim_{n\to 0,\beta\to\infty} \mathcal{A}_{i,n}^{(\beta)},$$

$$\mathcal{A}_{i,n}^{(\beta)} = \lim_{\beta\to\infty} \int \prod_{s=1}^{r} x_{s,i} \mathbb{E}_{\boldsymbol{c},\boldsymbol{\lambda}} \left[Z^{n-r} \prod_{s=1}^{r}\prod_{\mu=1}^{M} p_{y|z}(y_\mu|z_{s,\mu})^{\beta c_\mu} \prod_{i=1}^{N} e^{-\beta\lambda_i|x_{s,i}|} \delta(\boldsymbol{z}_s - A\boldsymbol{x}_s) \right] d^r\boldsymbol{x} d^r\boldsymbol{z}$$

$$\overset{\text{if } r\leq n\in\mathbb{N}}{=} \int \prod_{s=1}^{r} x_{s,i} \prod_{\mu=1}^{M} \mathbb{E}_{c_\mu} \left[\prod_{s=1}^{n} p_{y|z}(y_\mu|z_{s,\mu})^{\beta c_\mu} \right] \prod_{i=1}^{N} \mathbb{E}_{\lambda_i} \left[\prod_{s=1}^{n} e^{-\beta\lambda_i|x_{s,i}|} \right] \prod_{s=1}^{n} \delta(\boldsymbol{z}_s - A\boldsymbol{x}_s) d^n\boldsymbol{x} d^n\boldsymbol{z}$$

* 各因子が自明な高次元分布上の積分にすり替え（$(N+M)$ 次元 $\to$ $(M+N)n$ 次元）
* 多体問題の解の平均から、一体問題の平均へ
* $n \in \mathbb{N}$ での結果を書きくだし、$n \to 0$ へ外挿
  - レプリカ法／レプリカトリック

» リサンプリング統計量の生成関数としての自由エネルギー

* より一般的には、次の自由エネルギーがわかればよい

$$f(D,\lambda_0) \equiv \lim_{\substack{N,\beta\to\infty \\ n\to 0}} \frac{-1}{N\beta n} \log \Xi_n$$

$$\Xi_n = \int \prod_{\mu=1}^{M} \mathbb{E}_{c_\mu} \left[\prod_{s=1}^{n} p_{y|z}(y_\mu | z_{s,\mu})^{\beta c_\mu} \right] \prod_{i=1}^{N} \mathbb{E}_{\lambda_i} \left[\prod_{s=1}^{n} e^{-\beta\lambda_i |x_{s,i}|} \right] \prod_{s=1}^{n} \delta(\boldsymbol{z}_s - A\boldsymbol{x}_s) d^n\boldsymbol{x} d^n\boldsymbol{z}$$

* 取得データ D に依存
* 具体的な D の実現値に対して計算するのは難しい
* 自己平均性を期待する
  $f(D) = \mathbb{E}_D[f]$
  - 自己平均性は多くの平均場漸近論では正当化される場合が多い
  - とはいえ、これも簡単ではない（対数の平均）

» 典型評価のためのレプリカ

$$\mathbb{E}_D[f(D)] = \lim_{\substack{N,\beta\to\infty \\ n\to 0}} \frac{-1}{Nn\beta}\mathbb{E}[\log \Xi_n] = \lim_{\substack{N,\beta\to\infty \\ n,n'\to 0}} \frac{-1}{Nnn'\beta}\log \mathbb{E}\left[\Xi_n^{n'}\right],$$

$$\Xi_n^{n'} \overset{n'\in\mathbb{N}}{=} \int \prod_{\mu=1}^{M} q_{y|z}(y_\mu|z_{0,\mu})\mathbb{E}_{c_\mu}\left[\prod_{s=1}^{n}\prod_{k=1}^{n'} p_{y|z}(y_\mu|z_{sk,\mu})^{\beta c_\mu}\right]$$

$$\times \prod_{i=1}^{N} q_{x_0}(x_{0,i})\mathbb{E}_{\lambda_i}\left[\prod_{s=1}^{n}\prod_{k=1}^{n'} e^{-\beta\lambda_i|x_{sk,i}|}\right]\mathbb{E}_A\left[\prod_{s=1}^{n}\prod_{k=1}^{n'}\delta(\boldsymbol{z}_{sk} - A\boldsymbol{x}_{sk})\right] d^{nn'}\boldsymbol{x} d^{nn'}\boldsymbol{z} d\boldsymbol{x}_0 d\boldsymbol{z}_0 d\boldsymbol{y}$$

- 赤字部分が D 平均
- $n, n' \in \mathbb{N}$ での表式を計算し、$n, n' \to 0$ へ外挿
- A 平均以外はほぼ自明

» A 平均

* $A = USV^\top$ の回転不変性
  * $\Rightarrow$ $\tilde{\boldsymbol{z}}_{sk} \equiv U^\top \boldsymbol{z}_{sk}, \tilde{\boldsymbol{x}}_{sk} \equiv V^\top \boldsymbol{x}_{sk}$ は以下の拘束を満たす球面に一様分布
    $\tilde{\boldsymbol{z}}_{sk}^\top \tilde{\boldsymbol{z}}_{tl} = \boldsymbol{z}_{sk}^\top \boldsymbol{z}_{tl} = NQ_x^{(stkl)}, \quad \tilde{\boldsymbol{x}}_{sk}^\top \tilde{\boldsymbol{x}}_{tl} = \boldsymbol{x}_{sk}^\top \boldsymbol{x}_{tl} = MQ_z^{(stkl)},$
    $s, t = 0, 1, \ldots, n, \quad k, l = 1, \ldots, n'$

$$\mathbb{E}_A\left[\prod_{s=0}^{n}\prod_{k=1}^{n'}\delta(\boldsymbol{z}_{sk} - A\boldsymbol{x}_{sk})\right] = \int \prod_{\substack{s\le t\\ k\le l}} \delta(NQ_x^{(stkl)} - \boldsymbol{x}_{sk}^\top \boldsymbol{x}_{tl})\delta(MQ_z^{(stkl)} - \boldsymbol{z}_{sk}^\top \boldsymbol{z}_{tl})\mathbb{E}\left[\prod_{s=0}^{n}\delta(\tilde{\boldsymbol{z}}_s - S\tilde{\boldsymbol{x}}_s)\right] dQ_x dQ_z$$

» A 平均

* $A = USV^\top$ の回転不変性

  $\Rightarrow$ $\tilde{\boldsymbol{z}}_{sk} \equiv U^\top \boldsymbol{z}_{sk}, \tilde{\boldsymbol{x}}_{sk} \equiv V^\top \boldsymbol{x}_{sk}$ は以下の拘束を満たす球面に一様分布

  $\tilde{\boldsymbol{z}}_{sk}^\top \tilde{\boldsymbol{z}}_{tl} = \boldsymbol{z}_{sk}^\top \boldsymbol{z}_{tl} = NQ_x^{(stkl)}, \quad \tilde{\boldsymbol{x}}_{sk}^\top \tilde{\boldsymbol{x}}_{tl} = \boldsymbol{x}_{sk}^\top \boldsymbol{x}_{tl} = MQ_z^{(stkl)},$

  $s,t = 0,1,\ldots,n, \quad k,l = 1,\ldots,n'$

$$\mathbb{E}_A\left[\prod_{s=0}^{n}\prod_{k=1}^{n'}\delta(\boldsymbol{z}_{sk} - A\boldsymbol{x}_{sk})\right] = \int \prod_{\substack{s\le t\\ k\le l}} \delta(NQ_x^{(stkl)} - \boldsymbol{x}_{sk}^\top \boldsymbol{x}_{tl})\delta(MQ_z^{(stkl)} - \boldsymbol{z}_{sk}^\top \boldsymbol{z}_{tl})\mathbb{E}\left[\prod_{s=0}^{n}\delta(\tilde{\boldsymbol{z}}_s - S\tilde{\boldsymbol{x}}_s)\right] dQ_x dQ_z$$

* 行列積分による評価

$$\mathbb{E}\left[\prod_{s=0}^{n}\delta(\tilde{\boldsymbol{z}}_s - S\tilde{\boldsymbol{x}}_s)\right] = \frac{\int \prod_{s=0}^{n}\delta(\tilde{\boldsymbol{z}}_s - S\tilde{\boldsymbol{x}}_s)\prod_{\substack{s\le t\\ k\le l}}\delta(NQ_x^{(stkl)} - \tilde{\boldsymbol{x}}_{sk}^\top \tilde{\boldsymbol{x}}_{tl})\delta(MQ_z^{(stkl)} - \tilde{\boldsymbol{z}}_{sk}^\top \tilde{\boldsymbol{z}}_{tl}) d^{n+1}\tilde{\boldsymbol{x}} d^{n+1}\tilde{\boldsymbol{x}}}{\int \prod_{\substack{s\le t\\ k\le l}}\delta(NQ_x^{(stkl)} - \tilde{\boldsymbol{x}}_{sk}^\top \tilde{\boldsymbol{x}}_{tl})\delta(MQ_z^{(stkl)} - \tilde{\boldsymbol{z}}_{sk}^\top \tilde{\boldsymbol{z}}_{tl}) d^{nn'+1}\tilde{\boldsymbol{x}} d^{nn'+1}\tilde{\boldsymbol{x}}}$$

$$\to e^{N g_{\mathrm{G}}(Q_x, Q_z) - N g_{\mathrm{S}}(Q_x, Q_z)}$$

» 一般形

＊ $N \gg 1$ で各種積分を鞍点評価すると、一般的に以下の形になる

$$\frac{1}{N}\log \Xi_n^{n'} = \underset{Q_x, Q_z}{\mathrm{extr}} \left[g_{\mathrm{F}}(Q_x, Q_z) + g_{\mathrm{G}}(Q_x, Q_z) - g_{\mathrm{S}}(Q_x, Q_z) \right]$$

$$g_{\mathrm{F}} = \underset{\tilde{Q}_{1x}, \tilde{Q}_{1z}}{\mathrm{extr}} \left[\frac{1}{2}\mathrm{Tr}(\tilde{Q}_{1x} Q_x) + \frac{\alpha}{2}\mathrm{Tr}(\tilde{Q}_{1z} Q_z) + \log \int e^{-\frac{1}{2}\boldsymbol{x}^\top \tilde{Q}_{1x} \boldsymbol{x}} q_{x_0}(x_0) \mathbb{E}_\lambda \left[\prod_{s,k} e^{-\beta\lambda |x_{sk}|} \right] d^{nn'+1}x \right.$$

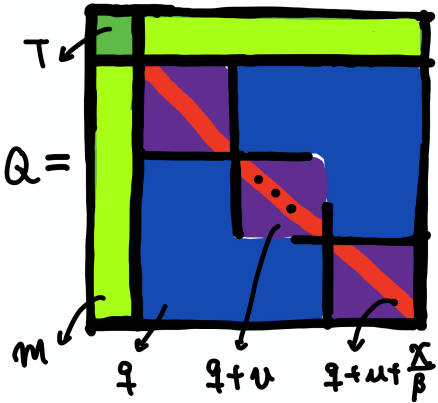
$$\left. + \alpha \log \int e^{-\frac{1}{2}\boldsymbol{z}^\top \tilde{Q}_{1z} \boldsymbol{z}} q_{y|z}(y|z_0) \mathbb{E}_c \left[\prod_{s,k} p_{y|z}(y|z_{sk})^{\beta c} \right] d^{nn'+1} z dy \right],$$

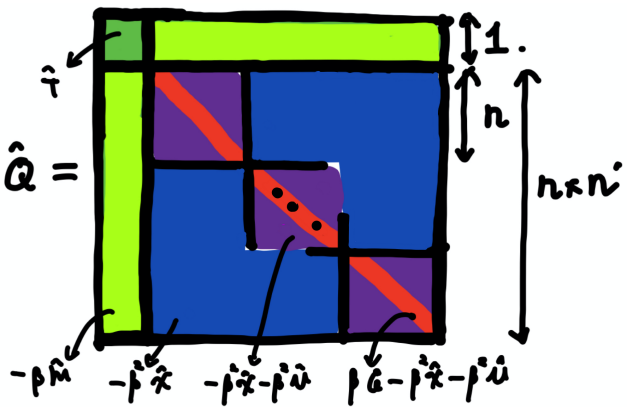
$$g_{\mathrm{G}} = \underset{\tilde{Q}_{2x}, \tilde{Q}_{2z}}{\mathrm{extr}} \left[\frac{1}{2}\mathrm{Tr}(\tilde{Q}_{2x} Q_x) + \frac{\alpha}{2}\mathrm{Tr}(\tilde{Q}_{2z} Q_z) + \frac{1}{2}\mathbb{E}_\gamma \left[\log\det(\tilde{Q}_{2x} + \gamma \tilde{Q}_{2z}) \right] \right],$$

$$g_{\mathrm{S}} = \frac{1}{2}\log\det Q_x + \frac{\alpha}{2}\log\det Q_z$$

※ $Q, \tilde{Q}$ はそれぞれ $(nn'+1) \times (nn'+1)$ の大きさの行列

» レプリカ対称仮定





- 同色の部分を同じ値に指定（レプリカ対称仮定）
- 形式的には 1-step replica symmetry breaking の破れ変数 $\to 0$ の構造

» RS 自由エネルギー

ちょっとスライドに書ききれない…

- arXiv:2001.02824 の appendix C の 1-RSB solution
- arXiv:2003.08670 の pp.25-26

» (マクロな) リサンプリング統計量の表現

平均場漸近論でのマクロなリサンプリング統計量

$$\frac{1}{N}\sum_{i=1}^{N}\phi\left(\mathbb{E}_{\boldsymbol{c},\boldsymbol{\gamma}}\left[\psi\left(\hat{x}_i(\boldsymbol{c},\boldsymbol{\gamma})\right)\right]\right) \overset{N\to\infty}{\to} \mathbb{E}_{\xi,x_0}\left[\phi\left(\mathbb{E}_{\eta,\lambda}\left[\psi\left(S_\lambda(\hat{m}_{1x}x_0+\sqrt{\hat{\chi}_{1x}}\xi+\sqrt{\hat{v}}_{1x}\eta,;\hat{Q}_{1x})\right)\right]\right)\right],$$

$$S_\lambda(a;b)=\frac{a-\lambda\,\mathrm{sign}(a)}{b}\mathbb{1}(|a|-\lambda),\quad \xi,\eta\sim\mathcal{N}(0,1),\ x_0\sim q_{x_0},\ \lambda\sim p_\lambda$$

$\hat{Q}_{1x},\hat{\chi}_{1x},\hat{m}_{1x},\hat{v}_{1x}$ は自由エネルギーの極値条件の解
ψ,ϕ は適当な関数

- 少数自由度 + 低次元積分での表現
  （リサンプリングを考慮した系での decoupling principle の表現）

» ここまで整理

* 準備：確率的推論の極限として推論問題を定式化
* リサンプリング平均, データ平均の階層性に応じて２度レプリカ法を使う
* 特徴量に関する平均はレプリカ系での行列積分で処理
  - $A^\top A$ の漸近固有値分布に関する平均処理となる
  - スケーリング極限が妥当になる範囲は…わからない
  - なんとなくランダム行列理論との接点を感じるところ
* 鞍点にレプリカ対称構造を仮定し、$n, n' \to 0$ 極限を計算
  - 大自由度の問題 $\to$ 少数自由度の問題

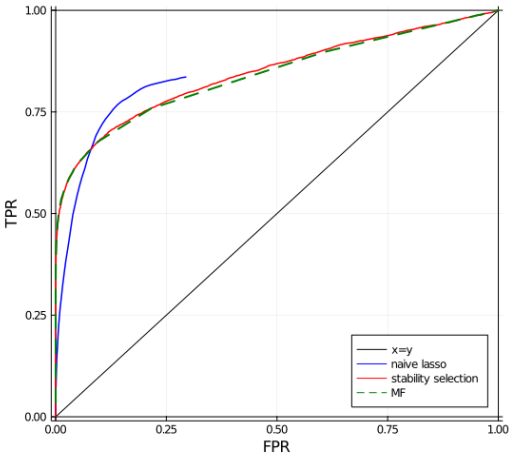
  ※ 形式的には 1RSB 解析の破れ変数 $\to 0$ に相当
* 鞍点の値と低次元積分でマクロなリサンプリング統計量が算出可能となる

» 平均場でのリサンプリング法の性質

* 興味の対象
  - FP を減らしたときに TP は保てるのか
  - 保てるのなら、それはどのような条件で
  - ブートストラップサンプル数にはどのような依存性があるか
* 手続き
  - λ_0 を通常の LASSO の leave-one-out 交差検証誤差最小化で決定
  - 漸近固有値分布は geometric ensemble で指定
    - $p(\lambda) \propto \lambda^{-1}, \lambda \in [a, b)$
    - $\kappa = b/a$: 条件数
    - ここでは $\kappa = 10$ で固定
  - 極値条件は反復計算で解く
  - このときの FPR, TPR のトレードオフを見る

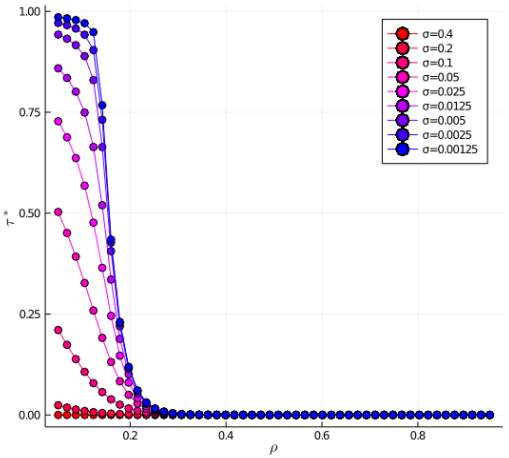
» ROC カーブ

- $\rho = 0.1, \sigma = 0.1, \alpha = 0.5, N = 512$
- 100 サンプル平均
- ナイーブな LASSO: 正則化パラメータを様々に変更して TPR, FPR をチェック
- $B = 5000$ の実験と理論は整合的
- ナイーブな LASSO は FPR $\to 0$ で TPR が非常に低い
- SS は FPR $\to 0$ で TPR $\to \tau^* > 0$ に出来る模様

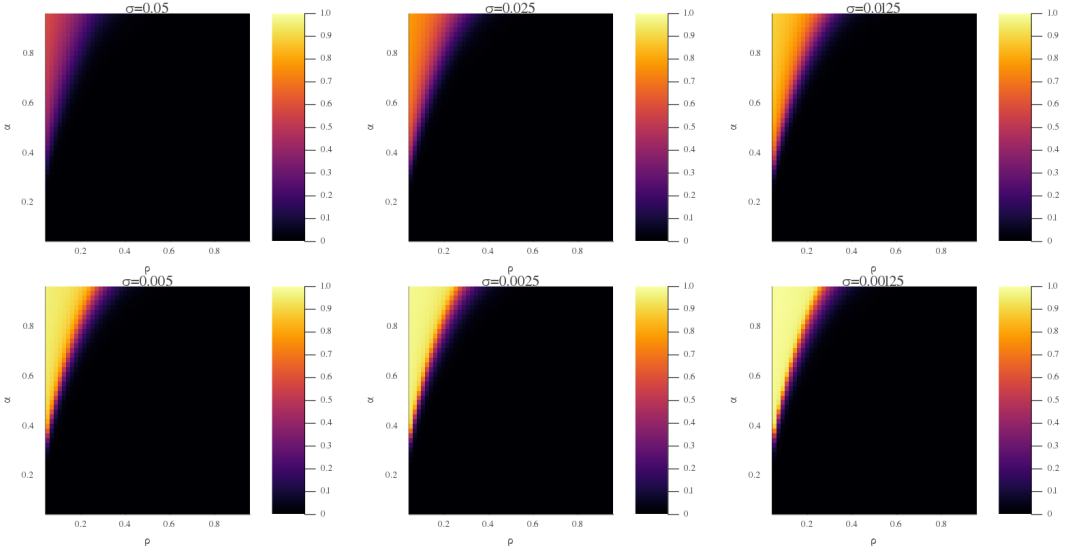


» τ^* の性質: α 一定

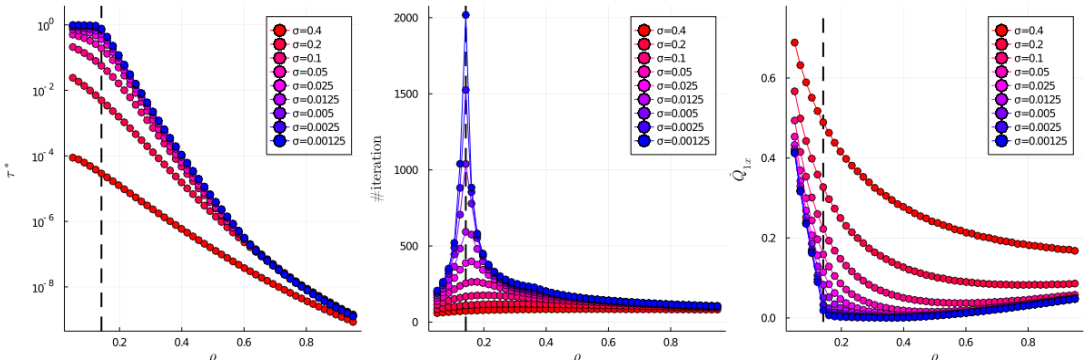
* $\alpha \simeq 0.7$, σ 依存性
* ある ρ を超えると得られる情報は急激に減衰



» τ^* の性質

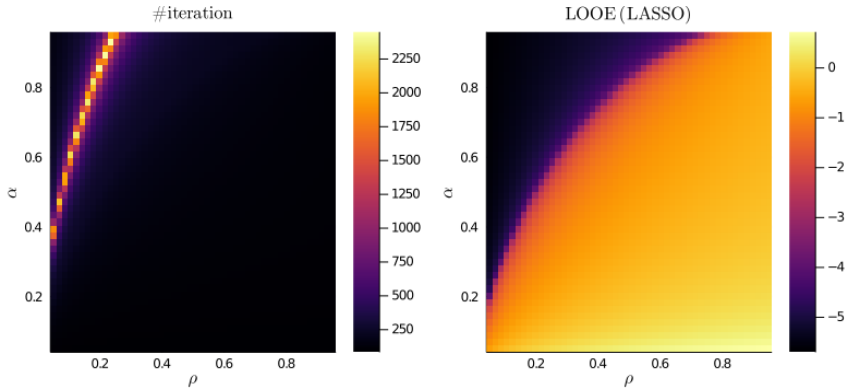

σ=0.05
σ=0.025
σ=0.0125
σ=0.005
σ=0.0025
σ=0.00125
α
ρ
0.2
0.4
0.6
0.8
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

» ゼロノイズでの相転移?



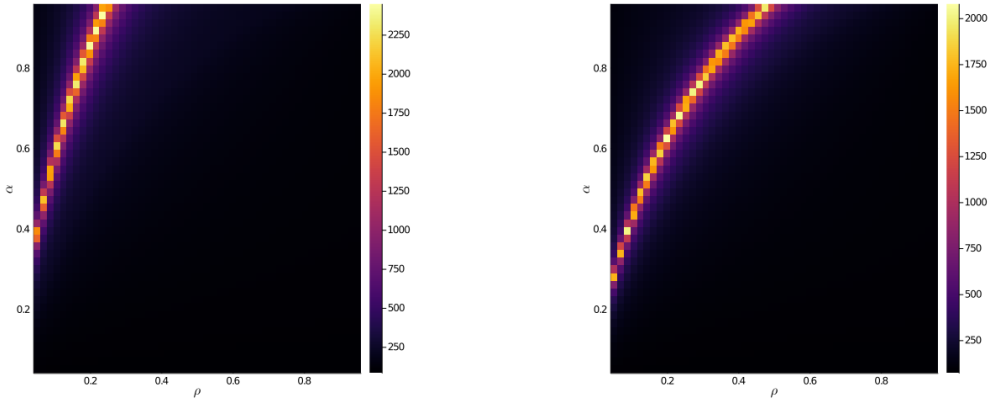
* 中央: 収束するのに要した反復回数
* $\sigma \to 0$ で固定点構造に特異性が出現しているようにみえる
* 詰めきれてないが、たぶん相転移になってるだろう

» ゼロノイズでの相転移?



* 左: 反復回数, 右: LASSO の最適な LOOE, $\sigma = 0.00125$
* 有限ノイズで $\tau^* > 0$ を達成する代わりに、ゼロノイズでの損をしている (ブートストラップサンプルの重複があるので…)

» 相境界のリサンプリングデータ数依存性



* 反復回数のリサンプリングデータサイズ依存性, $\sigma = 0.00125$
* 左: 大きさ M, 右: 大きさ $2M$
* ゼロノイズでの相境界は L_1 の完全復元限界に近づく

» ここまで整理

* SS を利用した場合 $\mathrm{FPR} \ll 1$ で TPR の値は保持できる
* ゼロノイズでは $\mathrm{TPR} = 1$ を保てる限界が存在し、そこよりも上側なら TPR は reasonablely well
* ただし、リサンプリングによる統計性獲得の代わりに検出限界については損をしている
* そこは多少リサンプリングデータのサイズを大きくとれたほうがよい
  - 計算量的な負荷は増える

イントロダクション

リサンプリング法の平均場漸近論

リサンプリング法の平均場での性質

まとめ

» まとめ

* 何を問題としていたのか
  - 高次元線形モデルの変数選択
  - 統計的手段を使うと素朴なスパース推定改善. そのときの exact な性能解析をしたい
* 解析のアイディア
  - ランダム問題, 平均場漸近論
  - 統計力学風の定式化
  - リサンプリング, データの平均のそれぞれに応じたレプリカ
* 解析結果
  - リサンプリングの統計性を用いると、false positive rate ゼロ極限で true positive rate> 0 達成可
  - ただし、スパースでない問題では難しい
  - 背後にゼロノイズでの相転移構造

» 不満点

* まったく数学的に厳密な導出ではない
  - 自由エネルギーについては適当な条件で厳密化可能か
  - ランダム行列理論の技術使えるか？
* 回転不変クラスは狭いように思われる
  - 実際は $\mathbb{E}_A[\prod_{s,k} \delta(\boldsymbol{z}_{sk} - A\boldsymbol{x}_{sk})]$ のスケーリング極限が同じものに行くものは全部同じ結果になるはず.
  - その限界は…

» 引用文献

[BKM14] Peter Bühlmann, Markus Kalisch, and Lukas Meier, *High-dimensional statistics with a view toward applications in biology*.

[BVDG11] Peter Bühlmann and Sara Van De Geer, *Statistics for high-dimensional data: methods, theory and applications*, Springer Science & Business Media, 2011.

[DMM09] David L Donoho, Arian Maleki, and Andrea Montanari, *Message-passing algorithms for compressed sensing*, Proceedings of the National Academy of Sciences **106** (2009), no. 45, 18914–18919.

[KWT09] Yoshiyuki Kabashima, Tadashi Wadayama, and Toshiyuki Tanaka, *A typical reconstruction limit for compressed sensing based on lp-norm minimization*, Journal of Statistical Mechanics: Theory and Experiment **2009** (2009), no. 09, L09003.

[MB10] Nicolai Meinshausen and Peter Bühlmann, *Stability selection*, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **72** (2010), no. 4, 417–473.

[Tib96] Robert Tibshirani, *Regression shrinkage and selection via the lasso*, Journal of the Royal Statistical Society: Series B (Methodological) **58** (1996), no. 1, 267–288.

[TK20a] Takashi Takahashi and Yoshiyuki Kabashima, *Macroscopic analysis of vector approximate message passing in a model mismatch setting*, arXiv preprint arXiv:2001.02824 (2020).

» 引用文献 (cont.)

[TK20b] _____, *Semi-analytic approximate stability selection for correlated data in generalized linear models*, arXiv preprint arXiv:2003.08670 (2020).