# Spectral Parameter Estimation of Random Matrix Models
# Cauchy雑音損失による次元復元

早瀬　友裕

Graduate School of Mathematical Sciences, The University of Tokyo

## Aim

- ノイズを含む サイズの大きい行列一個のサンプルから，情報を取り出したい.

例. 信号$A$にノイズ$Z$が加わっていると仮定:

$$Y = A + \sigma Z \tag{0.1}$$

1. $A$ : $p \times d$ 行列
2. $\sigma > 0$
3. $Z$ : $p \times d$　ランダム行列, 成分は独立でガウス分布$N(0, 1/d)$に従う

$A$　と $\sigma$をパラメータとみなして，$Y$のサンプル一つから推定したい

**Q**. サンプル数が少ない代わりに，サンプルサイズ$(p, d)$が大きいことを使った推定ができないか？

**A**. ランダム行列の特異値分布を用いた最適化問題に帰着

## Idea 1. ランダム行列の特異値分布の中心極限定理

$Y$の特異値（の二乗の）分布を考える:

$$\mu_{Y^TY} := \frac{1}{d}\sum_{k=1}^{d} \delta_{\lambda_k},$$

ここで$\lambda_1 \leq \ldots \lambda_d$は$Y^TY$の固有値. このとき, $Y$はランダム行列であったから当然$\mu_{Y^TY}$も確率的に揺らぐが, 実は次元$(d,p)$が大きければ, $\mu_{Y^TY}$は決定的な分布で近似できる(ランダム行列の特異値の中心極限定理[4]). この分布をパラメータ依存性を明示して$\mu^{\square}(A,\sigma)$と書くことにする.（$\square$は決定的であることのマーク，また実際は $\mu^{\square}(A,\sigma)$は$A$の特異値と$\sigma$にしかよらない.）

この事実に基づき, $\mu_{Y^TY}$と$\mu^{\square}(A,\sigma)$のKL-divergenceを最小化することにより, パラメータ$A,\sigma$を推定できないか, というアイディアが浮かぶ.

## Idea 2. Cauchy変換とマスター方程式

ところが$\mu_{Y^TY}$の分布関数の解析的な計算は難しいので, 代わりにCauchy変換を考えるとこれは数値的に計算できる：

**Definition 0.1.** 確率測度$\mu$のCauchy変換とは, $z \in \mathbb{C} \setminus \mathbb{R}$で以下のように定義される解析的な関数である：

$$G_\mu(z) := \int \frac{1}{z-t}\mu(dt).$$

$G$が計算できる理由は, これは以下のMaster方程式の解になっているからである:

$$G_\mu(z) = [z - R_{A,\sigma}(G_\mu(z))]^{-1}.$$

ここで $R_{A,\sigma}$は**R-変換**と呼ばれ, パラメータ$A,\sigma$に依存する$\mathbb{C}$（の領域）上の関数である. しかも，この方程式は反復法で数値的に解くことができる.[2] (*実際は$R_{A,\sigma}$が難しい関数になっていることがあり, その場合は元のモデルをうまく変形して, $R$を簡単にする手法が取られる[1])

## Idea 3. Cauchy雑音と確率的最急勾配法

従って, 経験分布 $\mu_{Y^TY}$とモデルの分布$\mu^{\square}(A,\sigma)$の コーシー変換を比較すればよい. ところが, 標本分布に対するCauchy変換は積分を経由せねばならず計算量がかかる. そこでCauchy変換と以下のCauchy分布 $P_\gamma$の関係に注目する：

$$P_\gamma(t) := \frac{1}{\pi}\frac{\gamma^2}{t^2+\gamma^2},$$

ここで$\gamma > 0$. Cauchy変換$G$の虚部をとると, Cauchy分布との畳み込みが得られる:

$$-\frac{1}{\pi}\mathrm{Im}G_\mu(x+i\gamma) = [P_\gamma * \mu](x), \; x \in \mathbb{R},$$

しかも, 任意に$\gamma > 0$を固定しても元の分布を識別可能である:

- $P_\gamma * \mu = P_\gamma * \nu$ if and only if $\mu = \nu$,
- 固定された確率測度$\nu$に対し, **Cauchy cross entropy**　を以下で定義すると,

$$H_\gamma(\nu,\mu) := \int -\log\left[P_\gamma * \mu(x)\right] P_\gamma * \nu(x) dx,$$

これが最小値をとるのは$\mu = \nu$と同値である.

$Y^TY$の固有値から一様に選んだ$\lambda_j$と, Cauchy分布から独立に生成した確率変数$T$(これをCauchy雑音と呼ぶ)を考えると

$$\lambda_j + T \sim P_\gamma * \mu_{Y^TY}$$

であるから, 積分を回避して, 結局以下の損失関数, Cauchy雑音損失の最小化問題に帰着される:

$$L_\gamma(A,\sigma) := \mathbb{E}_{X\sim\mu_{Y^TY}, T\sim P_\gamma}\left[\ell_\gamma(X+T+i\gamma)\right] := \mathbb{E}_{X\sim\mu_{Y^TY}, T\sim P_\gamma}\left[-\log\left(-\frac{1}{\pi}G_{\mu^{\square}(A,\sigma)}(X+T+i\gamma)\right)\right].$$

ここで経験分布とCauchy分布による積分は実際には実行されず，以下の確率的最急勾配法(確率的オンライン学習)で代用される. 注意として, Cauchy雑音とCauchy分布のスケールパラメータ$\gamma > 0$は学習前に固定されるハイパーパラメータである.

## Optimization Algorithm: Cauchy Noise SGD

**Cauchy Noise Stochastic Gradient Descent**

**Require** A $d \times d$ self-adjoint matrix $W$

$\lambda = (\lambda_1, \ldots, \lambda_d) \in \mathbb{R}^d \leftarrow$ eigenvalues of $W$

**for** $n \leftarrow 0, N-1$

  Compute learning rate $\mathrm{lr}^n$

  Shuffle $\lambda$

  **for** $j \leftarrow 1, d$

    **Generate samples** $t_k$ $(k = 1, \ldots, m)$ **from Cauchy(0,**$\gamma$**)**.

    $x_k \leftarrow \lambda_j - t_k$ $(k = 1, \ldots, m)$

    Compute $G_{W^{\square}(\vartheta^{nd+j})}(x_k + i\gamma)$ and $\nabla_\vartheta G_{W^{\square}(\vartheta)}(x_k + i\gamma)|_{\vartheta=\vartheta^{nd+j}}$ $(k = 1, \ldots, m)$.

    Calculate $\nabla_\vartheta \ell_\gamma(x_k, \vartheta)|_{\vartheta=\vartheta^{nd+j}}$ $(k = 1, \ldots, m)$.

    Update parameters by

$$\vartheta^{nd+j+1} \leftarrow \vartheta^{nd+j} - \mathrm{lr}^n \frac{1}{m}\sum_{k=1}^{m} \nabla_\vartheta \ell_\gamma(x_k, \vartheta)|_{\vartheta=\vartheta^{nd+j}}.$$

    $\vartheta^{nd+j+1} \leftarrow \Pi(\vartheta^{nd+j+1})$　　$\triangleright$ Project onto $\Theta$.

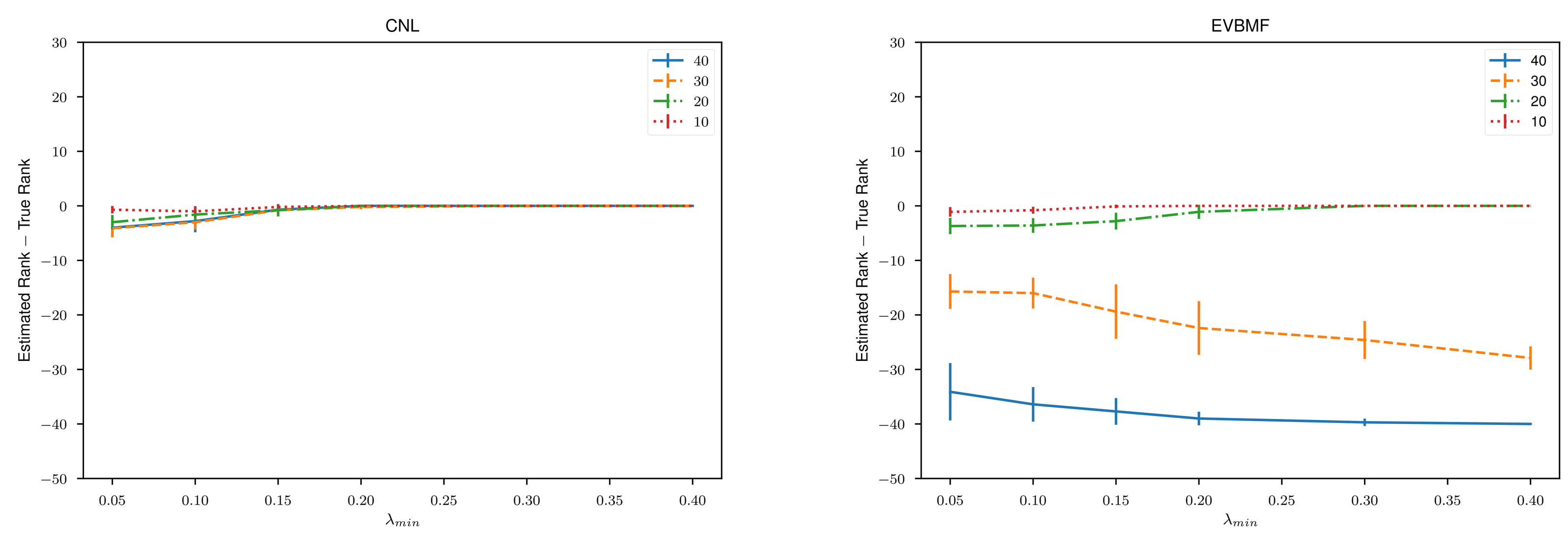
  **end for**

**end for**

**Ensure** $\vartheta^{Nd}$

Note that $\nabla_\vartheta \ell_\gamma(x,\vartheta) = -\mathrm{Im}\nabla_\vartheta G_{W^{\square}_\vartheta}(x+i\gamma)/\mathrm{Im}G_{W^{\square}_\vartheta}(x+i\gamma)$ and we compute $\nabla_\vartheta G_{W^{\square}_\vartheta}(x+i\gamma)$ using implicit differentiation

## Experiment 2: Dimensionality Recovery




Cauchy雑音損失(CNL)最小化により, 元の信号$A$のランクとノイズパワー$\sigma$を推定する. 行列サイズは$(p,d) = (100, 50)$とした. サンプルは$\mathrm{rank}A_{\mathrm{true}} = 10, 20, 30, 40$ と $\sigma_{\mathrm{true}} = 0.1$なるモデルから生成した. パラメータは$A$の特異値のベクトル$a$と $\sigma$である. 推定された$a$の成分の内、非ゼロの個数を数えればよい. しかし, SGDにより推定値がぶれるため, $L^1$正則化項を損失に加えて$a$の小さい成分を0に近づけた. 閾値$10^{-6}$で非ゼロを数えた. Cauchy雑音のハイパーパラメータ$\gamma > 0$は$10^{-4}$で固定した.

左が CNL最小化, 右が比較対象として, 経験変分ベイズ行列分解[3]による結果である. 横軸は$A_{\mathrm{true}}$の最小の非ゼロ特異値を表す. 縦軸は, 推定されたランクから, 真のランク$\mathrm{rank}A_{\mathrm{true}}$を差し引いたものを表す. したがって縦軸が0に近いものが推定精度が高いと見なされる.

CNL最小化はEVBMFと比べて, 元の信号が低ランクでなくとも精度よくランクを推定した.

## Links for ArXiv (left) and Github (right)

arXiv:1804.03154, github.com/ThayaFluss/cnl

## References

[1] S. T. Belinschi, T. Mai, and R. Speicher. Analytic subordination theory of operator-valued free additive convolution and the solution of a general random matrix problem. *J. Reine Angew. Math.*, 2013.

[2] J. W. Helton, R. R. Far, and R. Speicher. Operator-valued semicircular elements: solving a quadratic matrix equation with positivity constraints. *Int. Math. Res. Not.*, 2007, 2007.

[3] S. Nakajima, R. Tomioka, M. Sugiyama, and S. D. Babacan. Condition for perfect dimensionality recovery by variational Bayesian PCA. *J. of Mach. Learn. Res.*, 16(3757-3811):1, 2015.

[4] D. Voiculescu. Limit laws for random matrices and free products. *Invent. Math.*, 104(1):201–220, 1991.

LaTeX TikZposter