Cauchy noise loss for stochastic optimization of random matrix models via free deterministic equivalents

Tomohiro Hayase (Graduate School of Mathematical Sciences, The University of Tokyo)

arXiv:1804.03154, github.com/ThayaFluss/cnl

Aim

- To estimate parameters of random matrix models from a **single-shot** observation.

(Example of situation: analysis of covariance matrix, principal component analysis)
We introduce typical random matrix models as follows.

1. A *compound Wishart (CW) model* of type (p,d) is defined as a map

$$W_{\mathrm{CW}} \colon M_p(\mathbb{C})_{\mathrm{s.a.}} \to M_d(\mathbb{C})_{\mathrm{s.a.}};\ W_{\mathrm{CW}}(A) = Z^* A Z.$$

2. An *information-plus-noise (IPN) model* of type (p,d) is defined as a map

$$W_{\mathrm{IPN}} \colon M_{p,d}(\mathbb{C}) \times \mathbb{R} \to M_d(\mathbb{C})_{\mathrm{s.a.}};\ W_{\mathrm{IPN}}(A,\sigma) = (A+\sigma Z)^*(A+\sigma Z).$$

Note that Z is a $p \times d$ Gaussian random matrix with $\mathbb{E}[Z_{ij}] = 0, \mathbb{E}[Z_{ij}^2] = 1/d$.
Then our aim is to estimate the parameter ϑ_0 from a single-shot observation $W_{\vartheta_0}(\omega) = W(\vartheta_0)(\omega)$, $\omega \in \Omega$. Note that usual maximal likelihood estimation is not suitable, since we cannot use multiple observations.

Main Idea

If the matrix size if large, the Cauchy transform of the empirical spectral distribution (ESD) is close to a deterministic function. To apply **cross entropy minimization**, we focus on the γ-slice of Cauchy transform;

$$-\frac{1}{\pi}\mathrm{Im} G_{W_{\vartheta_0}}(x+i\gamma) = P_\gamma * \mathrm{ESD}(W_{\vartheta_0})(x),\ x \in \mathbb{R},$$

where $\gamma > 0$ and P_γ is a Poisson kernel, which is also the p.d.f. of Cauchy distribution of scale γ (denoted by Cauchy$(0,\gamma)$) ;

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

Note that the γ-slices have enough information to distinguish original probability measures; **for a fixed** $\gamma > 0$,

- $P_\gamma * \mu = P_\gamma * \nu$ if and only if $\mu = \nu$,
- for a fixed probability measure ν, the *Cauchy cross entropy*, defined by

$$H_\gamma(\nu,\mu) := \int -\log\left[P_\gamma * \mu(x)\right] P_\gamma * \nu(x) dx, \tag{0.1}$$

archives minimum if and only if $\mu = \nu$.

Free Deterministic Equivalents

Based on Speicher-Vargas [4], we define *free deterministic equivalent (FDE) models.* Its Cauchy transform approximates that of each original random matrix model. Let $(C_{ij}/\sqrt{d})_{i=1,\dots,p,j=1,\dots,d}$ be a standard $*$-free circular family in a C*-probability space.

1. The *FDECW model* is a map $W_{\mathrm{CW}}^{\square} \colon M_p(\mathbb{C}) \to M_d(\mathfrak{A})_{\mathrm{s.a.}}$ defined by

$$W_{\mathrm{CW}}^{\square}(A) = C^* A C.$$

2. The *FDEIPN model* is a map $W_{\mathrm{IPN}}^{\square} \colon M_{p,d}(\mathbb{C}) \times \mathbb{C} \to M_d(\mathfrak{A})_{\mathrm{s.a.}}$ defined by

$$W_{\mathrm{IPN}}^{\square}(A,\sigma) = (A+\sigma C)^*(A+\sigma C).$$

Iterative Methods

1. $G_{W_{\mathrm{CW}}^{\square}}$ is given by a limit of iteration of a contraction mapping.
2. $G_{W_{\mathrm{IPN}}^{\square}}$ is given by a limit of two nested loops of iteration of contraction mappings.

These contraction mappings consist of (matrix valued) R**-transform**, the **linearization trick**, and the **subordination**. See Helton-Far-Speicher[2] and Belinschi-Mai-Speicher[1] for more detail.

Cauchy Noise Loss

Instead of minimize$_{\vartheta \in \Theta} H_\gamma(\mu_{W_{\vartheta_0}^{\square}}, \mu_{W_{\vartheta}^{\square}})$, we try to minimize the empirical one;

$$\underset{\vartheta \in \Theta}{\mathrm{minimize}}\, H_\gamma(\mathrm{ESD}(W_{\vartheta_0}), \mu_{W_{\vartheta}^{\square}}).$$

To reduce the time complexity of computing the objective function, we introduce the *Cauchy noise loss*; for $\gamma > 0$ and $m \in \mathbb{N}$, we define

$$L_{\gamma,m}(\vartheta) := \frac{1}{dm}\sum_{j=1}^{d}\sum_{k=1}^{m} \ell_\gamma(\lambda_j - T_{j,k}, \vartheta), \tag{0.2}$$

where

$$\ell_\gamma(x,\vartheta) := -\log\left[-\frac{1}{\pi}\mathrm{Im}\, G_{W_{\vartheta}^{\square}}(x+i\gamma)\right], \tag{0.3}$$

$\lambda_1 \le \dots \le \lambda_d$ is the empirical eigenvalues of W_{ϑ_0}, and $T_{j,k}(j = 1,\dots,d, k = 1,\dots,m)$ are independent random variables distributed with the Cauchy distribution of scale γ.

Optimization Algorithm: Cauchy Noise SGD

Cauchy Noise Stochastic Gradient Descent

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

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

for $n \leftarrow 0, N-1$

  Compute learning rate lr^n

  Shuffle λ

  for $j \leftarrow 1, d$

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

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

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

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

    Update parameters by

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

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

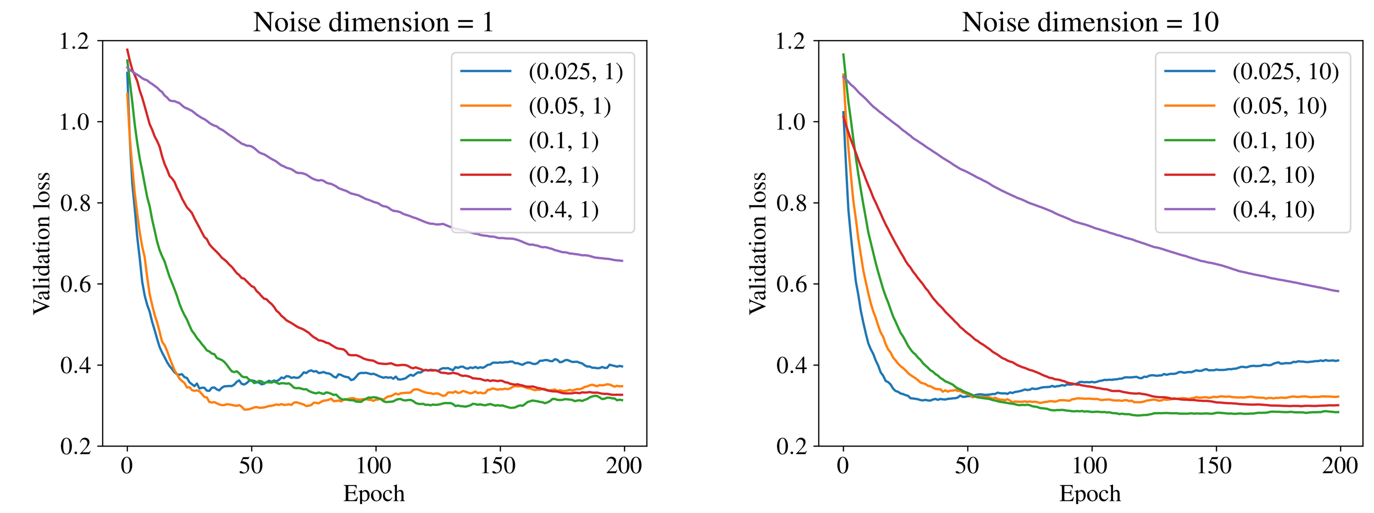
  end for

end for

Ensure ϑ^{Nd}

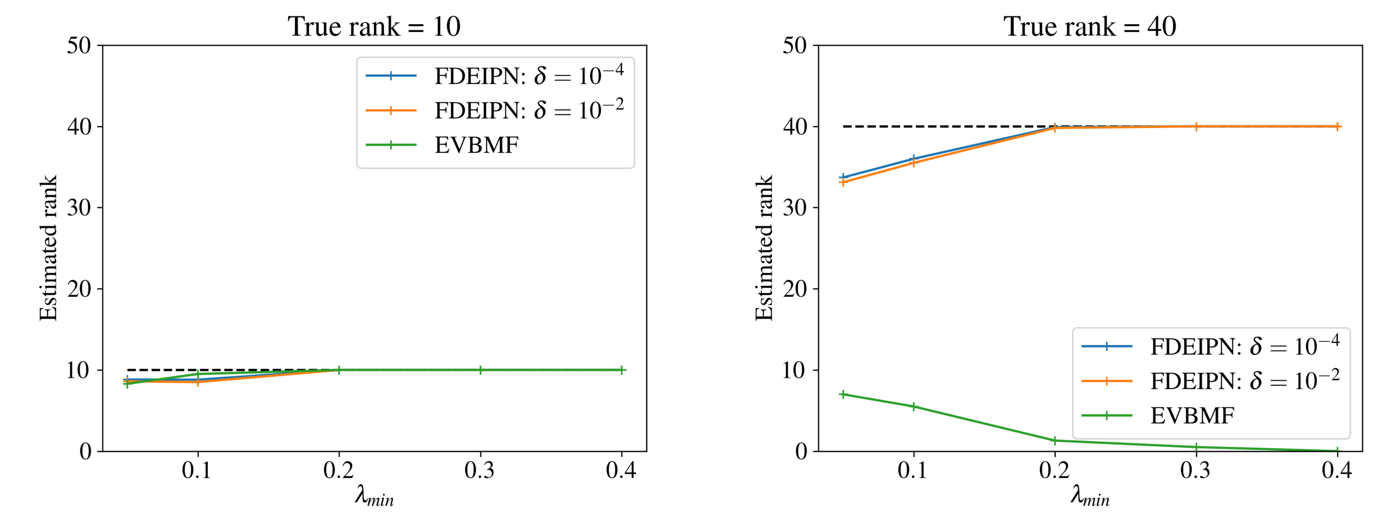
Note that $\nabla_\vartheta \ell_\gamma(x,\vartheta) = -\mathrm{Im}\nabla_\vartheta G_{W_{\vartheta}^{\square}}(x+i\gamma)/\mathrm{Im}G_{W_{\vartheta}^{\square}}(x+i\gamma)$ and we compute $\nabla_\vartheta G_{W_{\vartheta}^{\square}}(x+i\gamma)$ using implicit differentiation.

Experiment 1: Optimization of CW



The figure shows the optimization results of FDECW model of type $(p,d) = (50,50)$ under some values of (γ,m). The vertical axis indicates *the validation loss* which is a L^2-distance against the true parameter. **If** γ **is too small, the validation loss did not converge**.

Experiment 2: Rank Estimation



Rank estimation by FDEIPN with $(p,d) = (100,50)$ and the empirical variational Bayesian matrix factorization(EVBMF)[3]. The samples were generated from the model with rank$A_{\mathrm{true}} = 10, 40$ and $\sigma_{\mathrm{true}} = 0.1$. The horizontal axis $\lambda_{\min}$ represents minimum non-zero singular values of A_{true}. We set $(\gamma,m) = (p/d \times 0.2, 2)$. To shrink small parameters, we add a L^1 regularization term. True ranks were estimated even if they were not low.

Conclusion

We introduces optimization algorithm of random matrix models. It turned out that in experiments the key of the algorithm was the choice of γ.

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] R. Speicher and C. Vargas. Free deterministic equivalents, rectangular random matrix models, and operator-valued free probability theory. *Random Matrices Theory Appl.*, 1(02):1150008, 2012.