Matrix denoising via low-degree polynomials

Guilhem Semerjian

LPENS

28.01.2025 / IRS

based on J. Stat. Phys. 191, 139 (2024)

Outline

Outline

Inference problems

Inference : signal (S) ⟶ noisy channel ⟶ observations (Y)

Goal : recovering the signal from the observations

Bayesian setting : probabilistic model for the signal and the channel, known to the observer

all the useful information is in the posterior :

$$\mathbb{P}[S = s | Y = y] = \frac{\mathbb{P}[Y{=}y|S{=}s]\ \mathbb{P}[S{=}s]}{\mathbb{P}[Y{=}y]} \propto e^{-H_y(s)}$$

stat.mech. : observation $\Rightarrow$ quenched disorder

estimators : $\widehat{S}(Y)$, chosen to minimize (on average) some distance between the signal and its estimation

Matrix denoising

signal S, observations Y, $n \times n$ symmetric random matrices, $\widehat{S}(Y)$?

- multiplicative noise, $Y = \sqrt{S} Z \sqrt{S}$ $\quad$ S, Z independent
  covariance estimation : S/Y population/empirical covariance

- additive noise, $Y = S + Z$ $\quad$ S, Z independent
  simplification of matrix factorization $Y = XX^T + Z$ (or $Y = XF + Z$)
  X matrix $n \times r$ $\quad$ Z noise, e.g. Gaussian Orthogonal Ensemble

  well-understood in the low-rank regime $r = O(1)$
  [Rangan, Fletcher 12] [Lesieur, Krzakala, Zdeborová 17]
  [Lelarge, Miolane 19] [Barbier, Macris 19]
  or subextensive $r = o(n)$
  [Pourkamali, Barbier, Macris 23] [Barbier, Ko, Rahman 24]
  not so well for extensive ranks $r = \Theta(n)$
  [Maillard, Krzakala, Mézard, Zdeborová 22] [Barbier, Macris 22]
  [Camilli, Mézard 23] [Barbier, Camilli, Ko, Okajima 24]

Matrix denoising

large n limit, empirical spectral distribution of $S^{(n)} \to \mu_S$

$$\lim_{n\to\infty} \frac{1}{n}\mathbb{E}[\mathrm{Tr}((S^{(n)})^p)] = \int \mu_S(\mathrm{d}\lambda)\,\lambda^p \equiv \mu_{S,p}$$

idem for Z and Y with μ_Z and μ_Y

accuracy of an estimator $\widehat{S}(Y)$ in terms of the Mean Square Error

$$\begin{aligned}\mathrm{MSE}(\widehat{S}) &= \frac{1}{n}\sum_{i,j=1}^{n} \mathbb{E}[(S_{i,j} - \widehat{S}(Y)_{i,j})^2] \\ &= \frac{1}{n}\mathbb{E}[\mathrm{Tr}((S - \widehat{S}(Y))^2)]\ ,\end{aligned}$$

The BABP denoiser

estimator proposed in [Bun, Allez, Bouchaud, Potters 16]

try $\widehat{S}(Y)$ with the same eigenvectors as Y :

$$Y = \sum_{i=1}^{n} \lambda_i \, u_i u_i^T \qquad \widehat{S}(Y) = \sum_{i=1}^{n} \widehat{\lambda}_i \, u_i u_i^T \qquad S = \sum_{i=1}^{n} \zeta_i \, v_i v_i^T$$

minimizing the square error $\mathrm{Tr}((S - \widehat{S}(Y)^2)$ w.r.t. $\{\widehat{\lambda}_i\}$ yields :

$$\widehat{\lambda}_i = \sum_{j=1}^{n} \zeta_j (u_i^T v_j)^2$$

requires the knowledge of S (oracle estimator)

high-dimensional miracle : $\widehat{\lambda}_i \underset{n\to\infty}{\approx} \mathcal{D}_{\mathrm{BABP}}(\lambda_i)$

with $\mathcal{D}_{\mathrm{BABP}}$ a function that can be computed from μ_S, μ_Z, μ_Y

The BABP denoiser

$\widehat{S}(Y) = \mathcal{D}_{\mathrm{BABP}}(Y)$ with $\mathcal{D}_{\mathrm{BABP}} : \mathbb{R} \to \mathbb{R}$:

function acts on each eigenvalue

hence $\widehat{S}(OYO^T) = O\widehat{S}(Y)O^T$ for all $O \in \mathcal{O}_n$ (orthogonal group) :

equivariant function

BABP approach :

- not explicitly Bayesian (but oracle should be optimal among the equivariant)
- computation of the eigenvector overlaps with replicas (heuristic)

  rigorous in [Ledoit, Péché 11]

in the following : justification of the BABP estimator with explicit assumptions and elementary computations

done via HCIZ integrals in a special case

[Maillard, Krzakala, Mézard, Zdeborová 22] [Pourkamali, Barbier, Macris 23]

Outline

Bayesian estimation

more generic setting : (S, Y) pair of correlated r.v., $S \in \mathbb{R}^N$, $Y \in \mathbb{R}^M$

$\langle S, T \rangle = \sum_{i=1}^N S_i T_i$, $||S||^2 = \langle S, S \rangle$

$$\mathrm{MSE}(\widehat{S}) = \mathbb{E}[||S - \widehat{S}(Y)||^2] = \sum_{i=1}^{N} \mathbb{E}[(S_i - \widehat{S}_i(Y))^2] \ .$$

optimal estimator : $\widehat{S}^{\mathrm{opt}}(Y) = \mathbb{E}[S|Y]$, posterior mean

hard to compute in high dimensions

Low-degree polynomial method :

- for hypothesis testing [Hopkins, Steurer 17] [Kunisky, Wein, Bandeira 22]
- for estimation [Schramm, Wein 22] [Montanari, Wein 22]
- for constraint satisfaction problems [Bresler, Huang 22]

proofs of hardness results,

thought to emulate polynomial-time algorithms

Approximate Bayesian estimation

introduce a variational space with basic functions (e.g. polynomials)

$$\widehat{S}(Y) = \sum_{\beta \in \mathcal{A}} c_\beta b_\beta(Y) \ , \ \mathcal{A} : \text{finite set}, \ \ c : \text{variational parameters}$$

reduces to a quadratic optimization problem in a smaller space:

$$\begin{aligned} \mathrm{MSE}(\widehat{S}) &= \mathbb{E}[||S||^2] + \sum_{\beta,\beta' \in \mathcal{A}} c_\beta \mathcal{M}_{\beta,\beta'} c_{\beta'} - 2 \sum_{\beta \in \mathcal{A}} c_\beta \mathcal{R}_\beta \\ &= \mathbb{E}[||S||^2] + c^T \mathcal{M} c - 2 c^T \mathcal{R} \ , \end{aligned}$$

where $\mathcal{M}$ is a square matrix and $\mathcal{R}$ a vector, both of size $|\mathcal{A}|$:

$$\mathcal{M}_{\beta,\beta'} = \mathbb{E}[\langle b_\beta(Y), b_{\beta'}(Y) \rangle] \qquad \mathcal{R}_\beta = \mathbb{E}[\langle S, b_\beta(Y) \rangle]$$

Approximate Bayesian estimation

optimal MSE in this subspace:

$$\mathrm{MMSE}_{\mathcal{A}} = \mathbb{E}[||S||^2] + \inf_{c \in \mathbb{R}^{|\mathcal{A}|}} [c^T \mathcal{M} c - 2c^T \mathcal{R}]$$

reached for

$$\sum_{\beta' \in \mathcal{A}} \mathcal{M}_{\beta,\beta'} c_{\beta'} = \mathcal{R}_\beta \quad \forall \beta \in \mathcal{A}$$

$$\Leftrightarrow \; \mathcal{M} c = \mathcal{R}$$

$$\Leftrightarrow \; \mathbb{E}[\langle \widehat{S}(Y), b_\beta(Y) \rangle] = \mathbb{E}[\langle S, b_\beta(Y) \rangle] \quad \forall \beta \in \mathcal{A}$$

rk : to be compared with

$$\mathbb{E}[\widehat{S}^{\mathrm{opt}}(Y) \varphi(Y)] = \mathbb{E}[S \, \varphi(Y)] \quad \text{for all test functions } \varphi$$

$$\text{when } \widehat{S}^{\mathrm{opt}}(Y) = \mathbb{E}[S|Y]$$

low-degree polynomial method : $\{b_\beta\}$ = polynomials $\mathbb{R}^M \to \mathbb{R}^N$ of degree $\leq D$

Symmetries in approximate Bayesian estimation

how to choose the functions $\{b_\beta\}_{\beta \in \mathcal{A}}$?

the larger $\mathcal{A}$ the better $\mathrm{MMSE}_{\mathcal{A}}$, but more costly

$\Rightarrow$ Exploit the symmetries

G group acting through linear representations on $\mathbb{R}^N$ and $\mathbb{R}^M$

$g \cdot S \in \mathbb{R}^N$ the image of S under the transformation $g \in G$, $g \cdot Y$ idem

isometric assumption: $\langle g \cdot S, g \cdot T \rangle = \langle S, T \rangle$

definition of $f : \mathbb{R}^M \to \mathbb{R}^N$ equivariant (covariant) : $f(g \cdot Y) = g \cdot f(Y)$

G symmetry of the inference problem $\quad \Leftrightarrow \quad (g \cdot S, g \cdot Y) \stackrel{\mathrm{d}}{=} (S, Y)$

Consequences :

- $\widehat{S}^{\mathrm{opt}}(Y) = \mathbb{E}[S|Y]$ is equivariant
- no loss on $\mathrm{MMSE}_{\mathcal{A}}$ by taking the b_β equivariant

(Hunt-Stein lemma)

Outline

Matrix denoising with orthogonally invariant priors

back to the matrix denoising problem :

$Y^{(n)} = S^{(n)} + Z^{(n)}$ in $M_n^{\rm sym}(\mathbb{R})$

$G = \mathcal{O}_n = \{O \in M_n(\mathbb{R}) : OO^T = O^T O = \mathbb{1}_n\}$ orthogonal group

acts on $M_n^{\rm sym}(\mathbb{R})$ via conjugation : $O \cdot S = OSO^T$

assumptions : priors on S and Z orthogonally invariant

$O \cdot S \overset{\rm d}{=} S$ and $O \cdot Z \overset{\rm d}{=} Z$

hence $(O \cdot S, O \cdot Y) \overset{\rm d}{=} (S, Y)$: symmetry of the inference problem

S, Z independent and orthogonally invariant $\Rightarrow$ asymptotically free

$\mu_Y = \mu_S \boxplus \mu_Z$ (free additive convolution)

Matrix denoising with orthogonally invariant priors

best estimator of degree $\leq D$?

i.e. such that $\widehat{S}(Y)_{i,j}$ polynomial of degree at most D in $\{Y_{1,1}, Y_{1,2}, \dots, Y_{n,n}\}$

equivariant (under conjugation by orthogonal matrices) polynomials

are of the form $Y^p(\mathrm{Tr}(Y))^{q_1}(\mathrm{Tr}(Y^2))^{q_2} \dots (\mathrm{Tr}(Y^D))^{q_D}$

degree $= p + \sum_i iq_i \leq D$

consider first the "scalar" estimators $\widehat{S}(Y) = \sum_{p=0}^{D} c_p Y^p$

Matrix denoising with orthogonally invariant priors

$\widehat{S}(Y) = \sum_{p=0}^{D} c_p Y^p$ minimizes the MSE when $\mathcal{M}c = \mathcal{R}$, with :

$$\mathcal{M}_{p,p'} = \frac{1}{n}\mathbb{E}[\mathrm{Tr}(Y^{p+p'})] \qquad \mathcal{R}_p = \frac{1}{n}\mathbb{E}[\mathrm{Tr}(SY^p)] \qquad p,p' = 0,1,\dots,D$$

limit $n \to \infty$ of $\mathcal{M}$ and $\mathcal{R}$?

- $\mathcal{M}_{p,p'} \to \mathcal{M}^{(\infty)}_{p,p'} = \mu_{Y,p+p'}$ Hankel matrix, invertible $\forall D$
- with some free-probability computation (S, Z asymptotically free)

$$\mathcal{R}_p = \frac{1}{n}\mathbb{E}[\mathrm{Tr}(Y^{p+1})] - \frac{1}{n}\mathbb{E}[\mathrm{Tr}(Z(S+Z)^p)]$$

$$\to \mathcal{R}^{(\infty)}_p = \mu_{Y,p+1} - \sum_{m=1}^{p+1} \kappa_{Z,m} \sum_{\substack{j_1,\dots j_m \geq 0 \\ j_1+\dots+j_m = p+1-m}} \mu_{Y,j_1} \cdots \mu_{Y,j_m}$$

Matrix denoising with orthogonally invariant priors

given μ_S, μ_Z, $\mu_Y = \mu_S \boxplus \mu_Z$, for each finite D, large n limit of the optimal polynomial of degree $\leq D$, $\widehat{S}(Y) = \mathcal{D}^{(D)}(Y)$, obtained by solving a linear system of dimension $D+1$:

$$\int \mu_Y(\mathrm{d}\lambda)\mathcal{D}^{(D)}(\lambda)\lambda^p = \mathcal{R}_p^{(\infty)} \qquad \forall p \in \{0, 1, \dots, D\}$$

moreover $\mathcal{D}^{(D)} \to \mathcal{D}_{\mathrm{BABP}}$ as $D \to \infty$, because

$$\int \mu_Y(\mathrm{d}\lambda)\mathcal{D}_{\mathrm{BABP}}(\lambda)\lambda^p = \mathcal{R}_p^{(\infty)} \qquad \forall p \geq 0$$

hence $\mathcal{D}^{(D)} =$ orthogonal projection of $\mathcal{D}_{\mathrm{BABP}}$ on the subspace of polynomials of degree at most D, within $L^2(\mathbb{R}, \mu_Y)$

Matrix denoising with orthogonally invariant priors

the equivariant polynomials are linear combinations of

$$Y^p (\mathrm{Tr}(Y))^{q_1} (\mathrm{Tr}(Y^2))^{q_2} \dots (\mathrm{Tr}(Y^D))^{q_D}$$

we only considered $\widehat{S}(Y) = \sum_{p=0}^{D} c_p Y^p$

fortunately, the non-scalar terms are asymptotically irrelevant :

- when $n \to \infty$, $\mathrm{Tr}(Y^j)$ concentrates around $\mathbb{E}[\mathrm{Tr}(Y^j)]$
- more precisely, the (Gaussian) fluctuations of $\mathrm{Tr}(Y^j) - \mathbb{E}[\mathrm{Tr}(Y^j)]$ are not enough correlated with S to modify the MMSE
  (second-order free probability) [Mingo, Speicher 06]

concludes the justification of the optimality of BABP
(modulo the exchange of $n \to \infty$ and $D \to \infty$)

Matrix denoising with orthogonally invariant priors

Example : Wishart signal corrupted by Gaussian noise

- $S^{(n)} = \frac{1}{\sqrt{nr}} X^{(n)} (X^{(n)})^T - \frac{1}{\sqrt{\alpha}} \mathbb{1}_n$

  $X^{(n)}$ is an $n \times r$ matrix filled with i.i.d. $\mathcal{N}(0,1)$

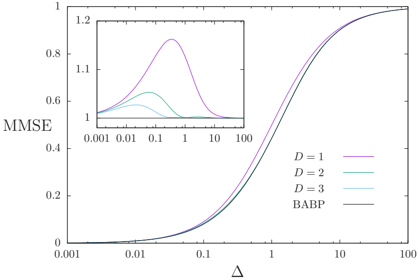
  $\alpha = n/r$

- $Z^{(n)} = \sqrt{\frac{\Delta}{n}} B$ with $B \sim$ GOE

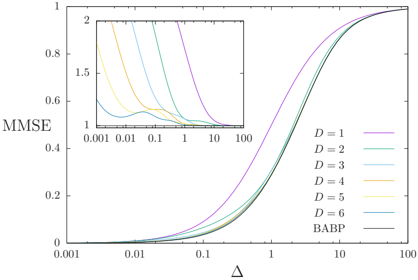
  $\{B_{i,j}\}_{i<j}$ i.i.d. $\mathcal{N}(0,1)$ $\qquad B_{i,j} = B_{j,i}$

  $\{B_{i,i}\}$ i.i.d. $\mathcal{N}(0,2)$

Matrix denoising with orthogonally invariant priors



$\alpha = 1$



$\alpha = 5$

larger α (smaller rank) requires larger D :

support of μ_Y made of two disjoint intervals,
$\mathcal{D}_{\text{BABP}}$ more difficult to approximate by polynomials

Outline

Beyond orthogonal invariance

what if the priors on S and Z are only asymptotically orthogonally invariant ?

many universality results in Random Matrix Theory :

- semi-circle law for Wigner matrices, not only GOE
- Marcenko-Pastur law for Wishart matrices (with X not necessarily Gaussian)
- eigenvectors delocalized and approximately isotropic
- freeness for Wigner matrices
- ...

are they strong enough to imply the universality of BABP as an optimal estimator ?

Beyond orthogonal invariance

Generalization of the example :

$$Y_{i,j} = \frac{1}{\sqrt{n}} \left(\frac{1}{\sqrt{r}} \sum_{\mu=1}^{r} X_{i,\mu} X_{j,\mu} + \sqrt{\Delta} B_{i,j} \right) = S_{i,j} + Z_{i,j}$$

$X_{i,\mu}$ i.i.d. $\mathbb{E}[X_{i,\mu}] = 0$, $\mathbb{E}[X_{i,\mu}^2] = 1$

$B_{i,j}$ i.i.d. $\mathbb{E}[B_{i,j}] = 0$, $\mathbb{E}[B_{i,j}^2] = 1$

with X and B not necessarily Gaussian

- non-universality in the law of B

  because $S_{i,j}$ and $Z_{i,j}$ are of the same order,

  scalar denoising problem not universal

- conjectured universality in the law of X (for finite D estimators)

Beyond orthogonal invariance

orthogonal invariance is broken, but permutation invariance remains :

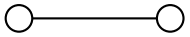
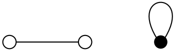
$$(\sigma \cdot S)_{i,j} = S_{\sigma(i),\sigma(j)} \qquad (\sigma \cdot S, \sigma \cdot Y) \stackrel{\mathrm{d}}{=} (S, Y)$$

equivariant polynomials (under permutations) indexed by multigraphs :

$$(b_G(Y))_{i,j} = \sum_{\substack{\phi \in [n]^V \\ \phi(v)=i, \phi(w)=j}} \prod_{e=\{a,b\}\in E} Y_{\phi(a),\phi(b)}$$

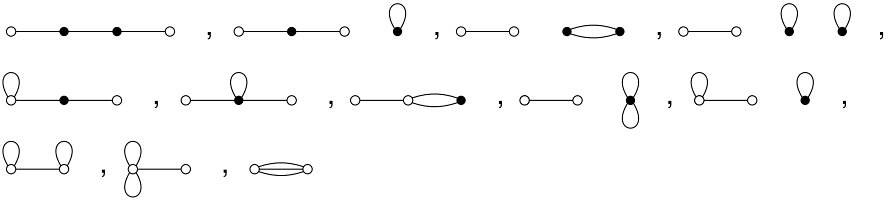
Examples :

cf. traffic distributions [Male 20]

-  $Y_{i,j}$
-  $(Y^2)_{i,j}$
-  $Y_{i,j} \, \mathrm{Tr}(Y)$
-  $Y_{i,j}(Y_{i,i} + Y_{j,j})$ not orthogonally equivariant

Beyond orthogonal invariance

Assuming inversion symmetry ($X \stackrel{d}{=} -X$, $B \stackrel{d}{=} -B$) these 4 were the only relevant of degree ≤ 2, and 12 more of degree 3 :



after some computations, large n limit of MMSE for polynomials of degree ≤ 3:

- is independent of the law of X
- is reached by $(\widehat{S}(Y))_{i,j} = c_1 Y_{i,j} + c_2 (Y^2)_{i,j} + c_3 (Y^3)_{i,j} + c_4 Y_{i,j}^3$
- if the noise is Gaussian, $c_4 = 0$, one finds back $\mathcal{D}^{(3)}$

Conclusions

- Low-degree polynomials versatile approach when a direct computation is not possible

  Believed to capture polynomial-time algorithms

- universality conjecture in the law of $X_{i,\mu}$ for $S = XX^T$:
  - strong version (optimal estimators) wrong for some laws and (α, Δ)
    - exponential-time algorithm better than BABP [Camilli, Mézard 23]
    - prior on X relevant in the sub-extensive rank regime [Pourkamali, Barbier, Macris 23] [Barbier, Ko, Rahman 24]
    - violates an information-theoretic bound [Barbier, Camilli, Ko, Okajima 24]
  - weak version (estimators with D finite) still open

  $\Rightarrow$ hard phases in the (α, Δ) phase diagram