Non-Asymptotic Analysis of Data Augmentation for Precision Matrix Estimation

Paper

NEURAL INFORMATION PROCESSING SYSTEMS

Lucas Morisset[1,2], Adrien Hardy[1], Alain Durmus[2]

[1] Qube Research and Technologies, [2] Ecole Polytechnique

Introduction

High-dimensional covariance matrices estimation

In high-dimension (number of feature d comparable to samples n), the sample covariance $C_X = n^{-1}XX^\top$, where $X \in \mathbb{R}^{n \times d}$, is a noisy estimate of the population covariance Σ.

Random matrix theory explains this: as $d, n \to \infty$ with $d/n \to \gamma$, the eigenvalue distribution of C_X converge to the Marchenko-Pastur distribution, which differs from the eigenvalues distribution of Σ.

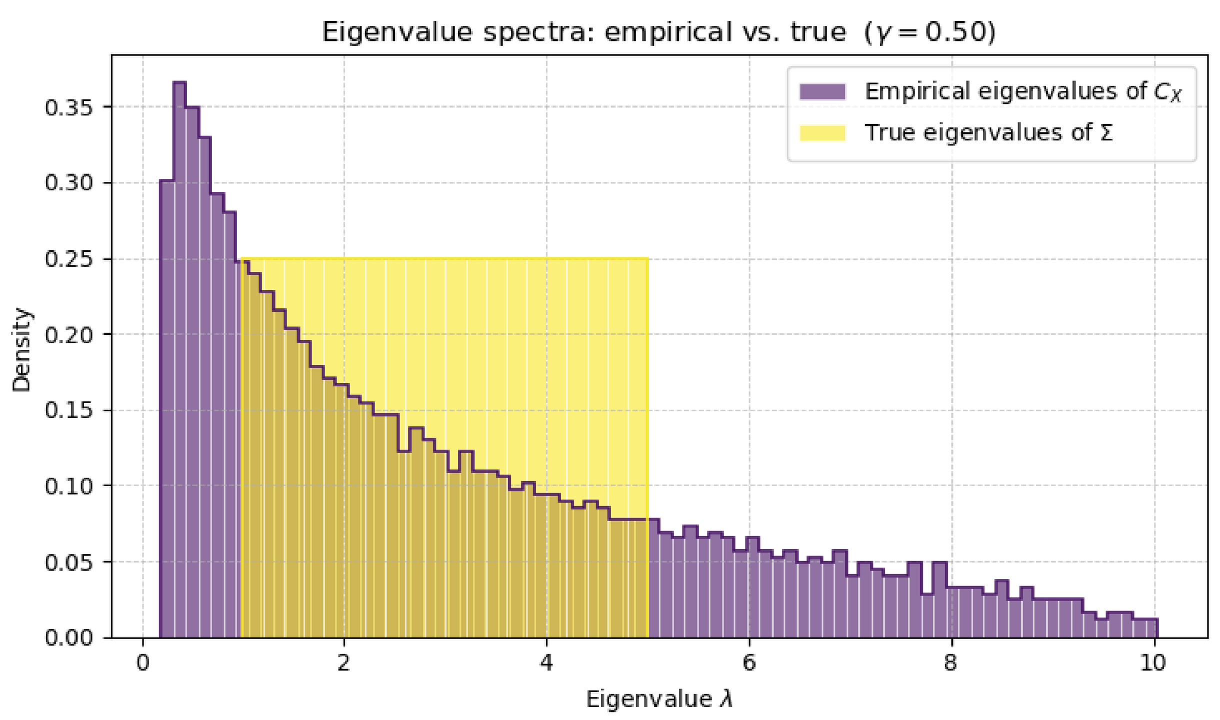

Figure 1 - $\Sigma = Diag\left((\lambda_i)_{i=1}^{2000}\right)$, where $\lambda_i = 1 + i\,/\,500$, and $C_X = n^{-1}XX^\top$, for $X_{i,:} \sim N(0, \Sigma)$.

Shrinkage estimators

To alleviate the high-dimensional effects, and in a scarse data regime, practitioners have developped so-called shrinkage estimators, which involves a combinaison of a noisy (high-variance) estimate, typically C_X, with a stable (low variance, high-bias) target $\mathbf{T}$, of the form,

$$S(\alpha) = (1-\alpha)\, C_X + \alpha \mathbf{T}, \quad \text{or} \quad S(\alpha) = C_X + \alpha \mathbf{T}.$$

Common choices for $\mathbf{T}$ include the identity matrix, or $\mathbf{T} = \mathrm{diag}(\mathrm{Cx})$.

Data Augmentation as a natural way to shrink your covariance

Consider an augmented dataset $X \sqcup G$, where G consists of artificial data. Assuming X and G are both centered, we rewrite the augmented covariance,

$$C_{X \sqcup G} = (1-\alpha)\, C_X + \alpha\, C_G, \qquad \text{with} \quad \alpha = |G|/(|X| + |G|).$$

Augmenting the dataset yields a shrinkage-like estimator, where $\mathbf{T} = C_G$.

Given a data augmentation scheme, can we optimize the induced regularization to produce more robust estimates of the covariance and precision matrices in the high-dimensional regime ?

Optimal shrinkage: Non Augmented case

Methodology

We consider a Ridge-like estimator of the precision matrix, define for $\lambda \geq 0$:

$$R_X(\lambda) = (C_X + \lambda\, I_d)^+, \qquad \mathcal{E}_X(\lambda) = d^{-1}\| R_X(\lambda) - \Sigma^{-1} \|_F$$

We estimate $\mathcal{E}_X(\lambda)$ up to an additive constant, and minimize our estimator w.r.t. λ. To this end, we define,

$$\hat{\mathcal{E}}_X(\lambda) = \frac{1}{d}\left(\mathrm{tr}\left(R_X(\lambda)^2\right) - \frac{2(1-\gamma_n)}{\lambda}\mathrm{tr}\left(R_X(0)\right) + \frac{2}{\lambda\,\rho(\lambda)}\,\mathrm{tr}\left(R_X(\lambda)\right)\right),$$

$$\rho(\lambda) = \frac{1}{1 - \gamma_n + {}^{\lambda}/_{n}\,\mathrm{tr}(R_X(\lambda))}, \qquad \gamma_n = d/n.$$

Then,

Meta-Theorem 1:

Assuming the samples of X are σ sub-Gaussian, we have for all $\lambda \geq 0$,

$$\left|\hat{\mathcal{E}}_X(\lambda) - \mathcal{E}_X(\lambda) + \frac{1}{d}\,tr(\Sigma^2)\right| \leq t + O\left(\frac{\sigma^2\sqrt{d}\lambda_{max}(\Sigma)^3}{n\,\eta^7} + \frac{1}{\eta^3 nd}\right),$$

where $\eta = \min\{\lambda, \lambda_{min}(\Sigma)\}$, and with probability $\geq 1 - \exp(-c\eta^4\sigma^2 n^2 t^2)$.

Numerical results on MNIST & CIFAR10

We estimate Σ using all the data (full curves) and simulate the high dimensional scenario by keeping only $n = d/\gamma$ samples (dashed curved).

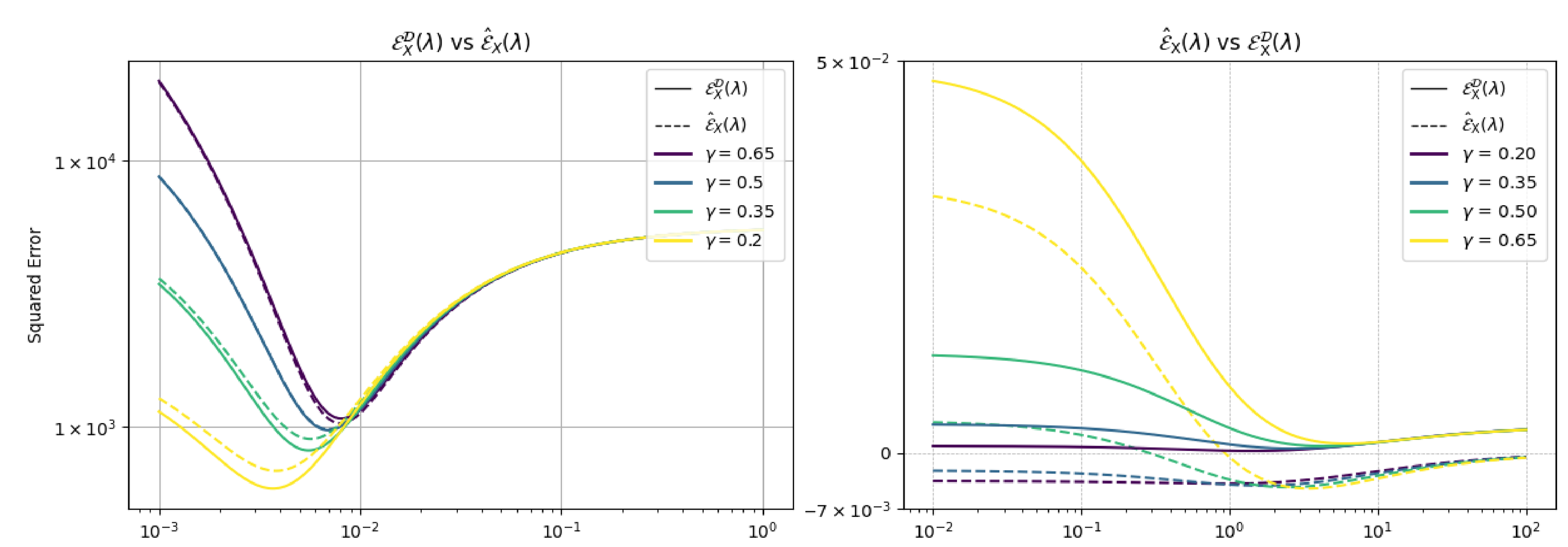

Figure 2 – Simulation of theorem 1 on MNIST (left) and CIFAR10 (right).

Optimal shrinkage: Augmented case

We consider the augmented estimator, and its error,

$$R_{Aug}(\lambda) = (C_{X \sqcup G} + \lambda\, I_d)^+, \qquad \mathcal{E}_{Aug}(\lambda) = d^{-1}\| R_{Aug}(\lambda) - \Sigma^{-1} \|_F.$$

Where the artificial dataset G is obtained by either

- Randomly transforming the true samples in X.
- Sampling from a complex generative model, fitted on X.

We make a sequence of assumption on the distribution of the artificial dataset (fully detailed in the paper):

- Σ is well conditionned.
- Samples of X are sub-Gaussian, and of the ones of G are sub-gaussian conditionally on X.
- The distribution of the artificial data is stable under small perturbation of X, and stable under removal of one of the samples of X.
- The data augmentation scheme can be sampled conditionally to X.

Then, we provide a function $\hat{\mathcal{E}}_{Aug}(\lambda)$ computable up to an additive constant, such that,

Meta-Theorem 2:

Under the previous set of assumptions, we have for all $\lambda \geq 0$,

$$\left|\hat{\mathcal{E}}_{Aug}(\lambda) - \mathcal{E}_{Aug}(\lambda)\right| \leq t + O\left(\frac{1}{\eta^9\sqrt{n}} + \frac{\lambda_{max}(\Sigma)^2 \|\,[E[C_G], \Sigma]\,\|_F}{\sqrt{n}\,\eta^2}\right)$$

where $\eta = \min\{\lambda, \lambda_{min}(\Sigma)\}$, and with probability going to 1 as $n \to +\infty$.

Numerical results on MNIST & CIFAR10

We consider two data augmentation satisfying the previous assumption:

- A Gaussian noise injection, $x \mapsto x + \sigma\,\varepsilon$, where $\varepsilon \sim N(0,1)$.
- A Gaussian mixture model fitted on X using the EM-algorithm.

We optimize the data augmentation over $\alpha = |G|/(|X| + |G|)$.

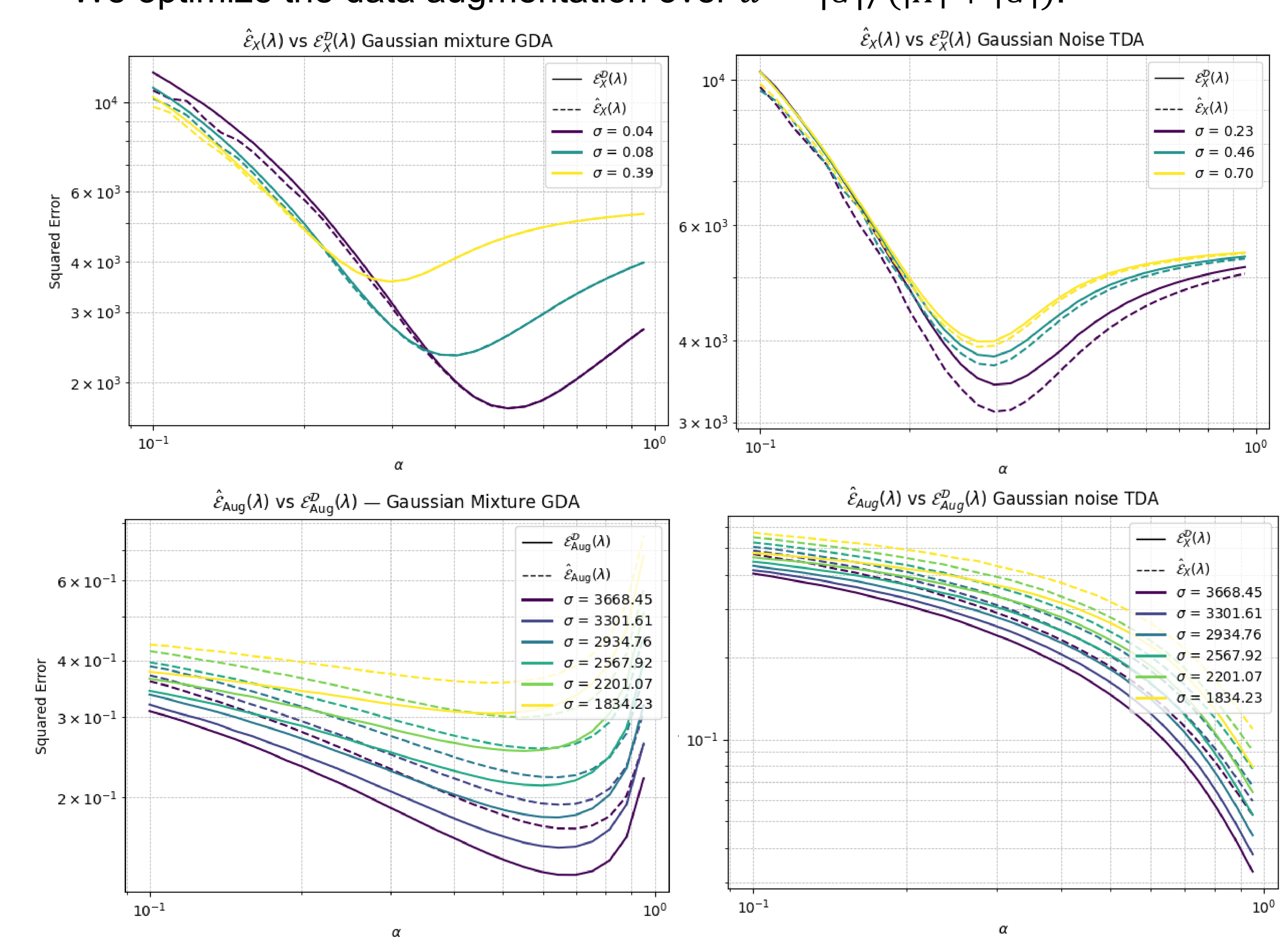

Figure 2 – Simulation of theorem 2 on MNIST (up) and CIFAR10 (down), for the Gaussian noise injection (right) and Gaussian mixture model (left)