About the slightly modified MP-PCA

Jonas L. Olesen and Sune N. Jespersen

December 15, 2020

This algorithm is a slightly improved version of the algorithm suggested by Veraart et al, see NeuroImage, Volume 142, 2016, Pages 394–406, `https://doi.org/10.1016/j.neuroimage.2016.08.016`.

1 Notation

MP theorem Consider a M x N real random matrix X with independent identically distributed elements of zero mean and variance σ^2. In the limit M, $N \to \infty$ with $M/N \equiv \gamma$ fixed, the eigenvalues λ of M x M matrix

$$H = \frac{1}{N} X X^T \tag{1}$$

are distributed according to the MP law.

Restate for symmetry in M and N, assuming it still holds approximately for finite sizes: To get rid of special status of the number N, consider instead the eigenvalues x of matrix

$$H = X X^T \tag{2}$$

This multiplies all eigenvalues by N ($N\lambda = x$), so they have pdf $p(x)$

$$p(x) = \frac{\sqrt{(x - x_-)(x_+ - x)}}{2\pi\sigma^2 M x} \tag{3}$$

with

$$x_\pm = N\sigma^2(1 \pm \sqrt{\gamma})^2 = \sigma^2(\sqrt{N} \pm \sqrt{M})^2 \tag{4}$$

This describes the full distribution if $M < N$. If $M > N$, there is a fraction of $(M-N)/M = 1-1/\gamma$ eigenvalues with value zero. If we remove those from the distribution, the remaining part of the distribution must be re-scaled by a factor $1/(1/\gamma) = M/N$:

$$p(x) = \frac{M}{N}\frac{\sqrt{(x-x_-)(x_+-x)}}{2\pi\sigma^2 Mx} = \frac{\sqrt{(x-x_-)(x_+-x)}}{2\pi\sigma^2 Nx} \tag{5}$$

$x_\pm$ are unchanged. Observe that the only difference between the cases $M > N$ and $M < N$ is whether we divide by M or N: In both cases, we divide by the smallest of the two. For brevity, we will refer to this MP derived distribution, as the MPX distribution.

Let's denote $M' = \min(M, N)$ and $N' = \max(M, N)$. Following Veraart et al, we make use of two properties

$$x_+ - x_- = 4\sigma^2\sqrt{M'N'} \tag{6}$$

$$\langle x\rangle = \int_{x_-}^{x_+} x dv(x) = \frac{1}{2\pi\sigma^2 M'}\frac{\pi}{8}(x_+ - x_-)^2 = N'\sigma^2 \tag{7}$$

Question: How does this change when there are p signal components?

Below is a simulation with $M = 100$, $N = 200$ and $p = 10$. SNR = 1e3 and $x_+ \simeq 582.8$ and lowest signal singular value 5e3. Repeated 1e4 times. Clearly, the best distribution is the MPX distribution with $M \to M - p$ and $N \to N - p$. This is consistent with other choices of M, N and p. The tail at the end which none of the distributions fit diminishes as matrix size increases.

To estimate the number of signal components p, we compute the variance in 2 different ways corresponding to equations (6) and (7):

$$\sigma_1^2 = \frac{1}{M'N'}\sum_{i=1}^{M'} x_i \tag{8}$$

and

$$\sigma_2^2 = \frac{x_1 - x_{M'}}{4'}\sqrt{\frac{1}{N'M'}} \tag{9}$$

When we have signal components, Veraart et al modify M and N, and the sums to only run over the $M' - p$ noise eigenvalues, and increase p from zero until $\sigma_2^2 \le \sigma_1^2$. Now the question is what to use for N' and M' when we have signal components? The original theory and method would have $M' \to M' - p$ and N' unchanged, whereas the results above would suggest also $N' \to N' - p$. These possibilities correspond to respectively

$$\frac{1}{(M'-p)N'}\sum_{i=p+1}^{M'} x_i \ge \frac{x_{p+1} - x_{M'}}{4}\frac{1}{\sqrt{N'(M'-p)}} \tag{10}$$

and

$$\frac{1}{(M'-p)(N'-p)}\sum_{i=p+1}^{M'} x_i > \frac{x_{p+1}-x_{M'}}{4}\frac{1}{\sqrt{(N'-p)(M'-p)}} \tag{11}$$

Oddly, it turns out that the best criteria statistically is none of the above, but instead

$$\frac{1}{(M'-p)(N'-p)}\sum_{i=p+1}^{M'} x_i \geq \frac{x_{p+1}-x_{M'}}{4}\frac{1}{\sqrt{N'M'}} \tag{12}$$

In the above example, these 3 estimates return p correctly in 6888, 8914 and 9996 cases respectively, out of 10000 repetitions.

We can estimate the variance of the underlying noise. This turns out to be most robustly estimated (different from Veraart et al) as

$$\sigma^2 = \frac{1}{(M'-p)(N'-p)}\sum_{i=p+1}^{M'} x_i \tag{13}$$

consistent with the above.

Final question pertains to how the residuals $X - X_d$, where X_d is the denoised matrix, are distributed. According to theory, they should be gaussian. Their variance is not σ^2, but $(M-p')/M'\sigma^2$ because some of the variability goes to perturb the signal singular values. However, it turns out that by far the best estimate of their variance is instead $(M'-p)/M'(N'-p)/N'\sigma^2$. This is illustrated in the next figure.

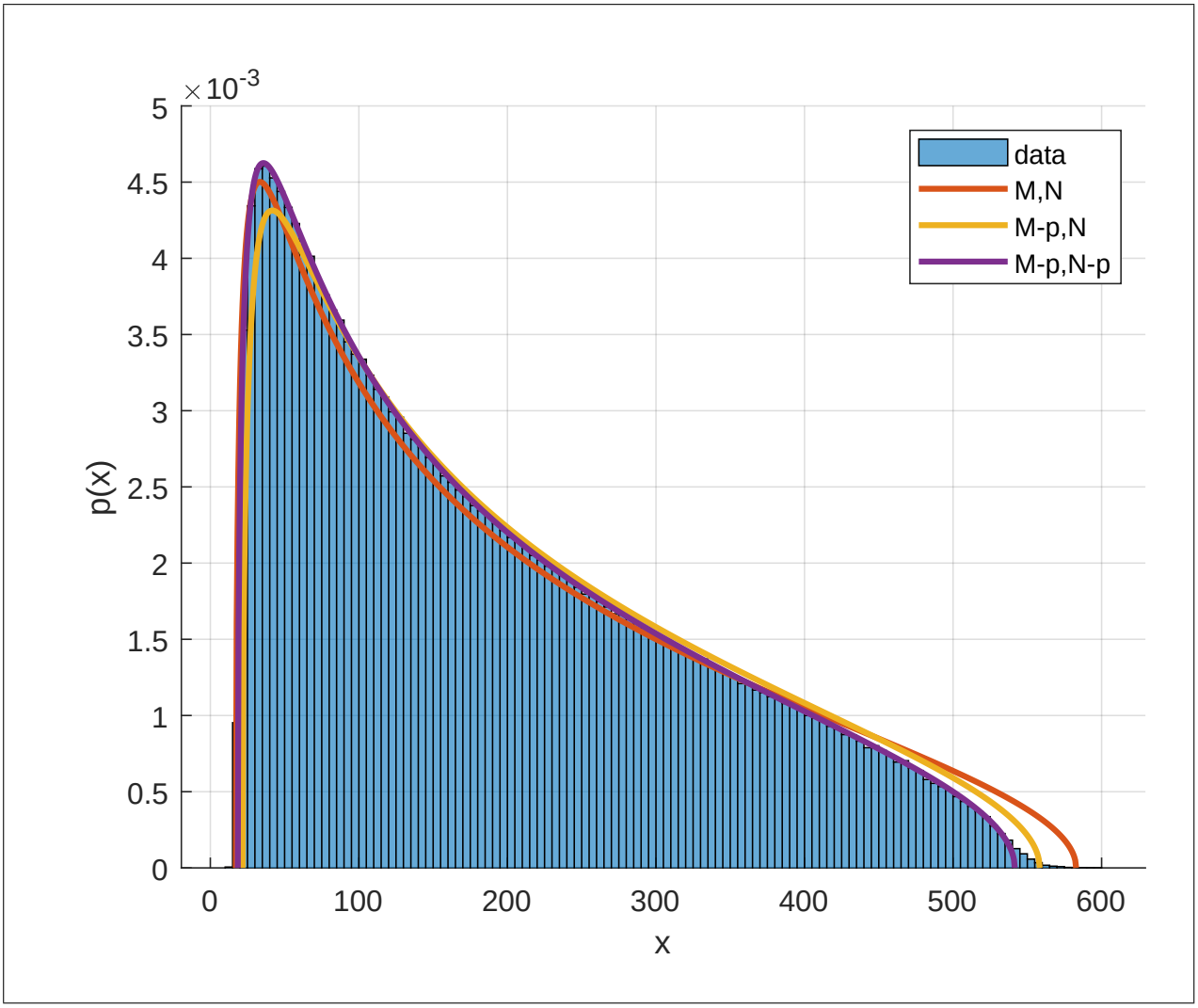

Figure 1: The distribution of $M - p$ smallest eigenvalues.

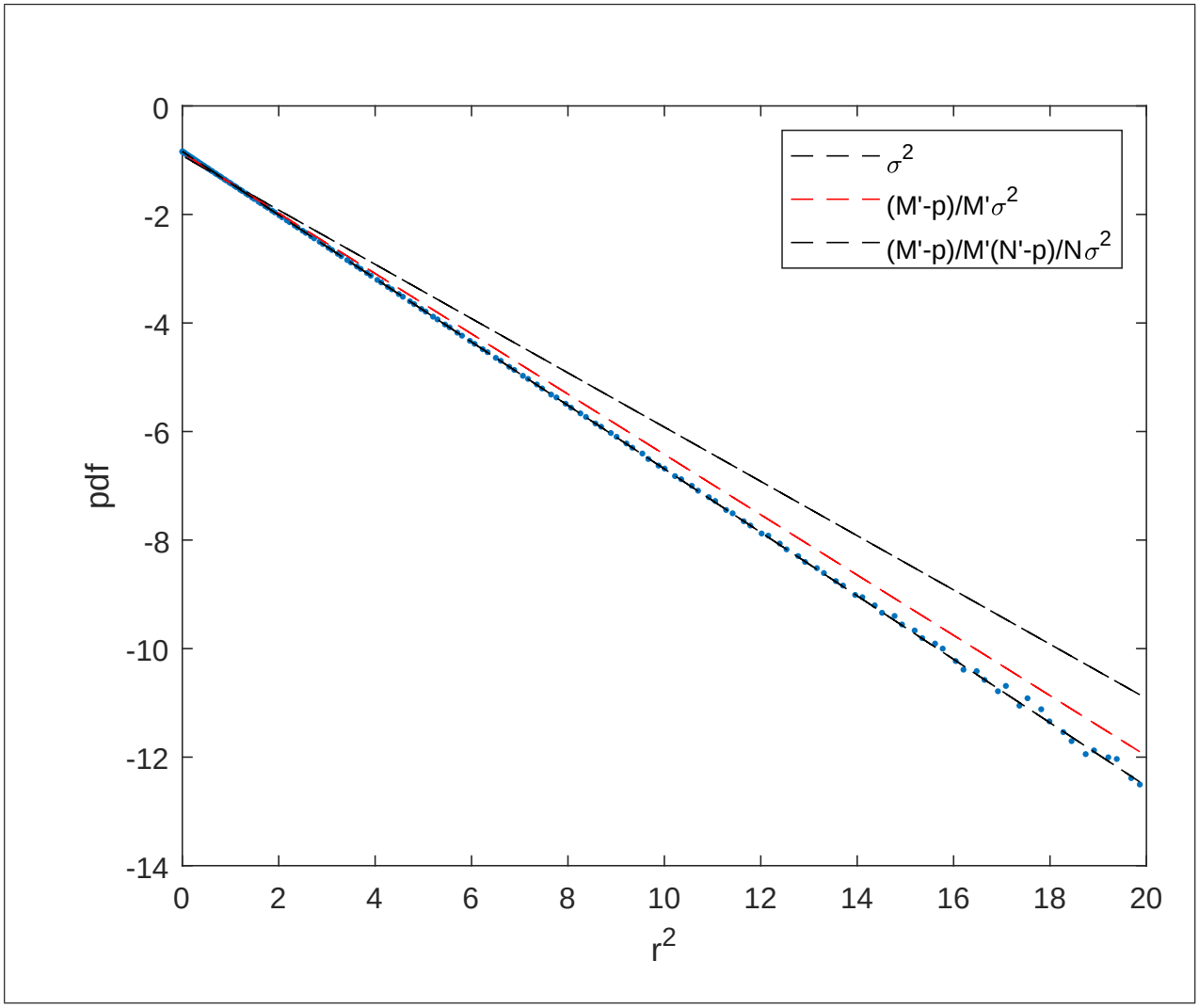

Figure 2: The distribution of residuals, $X - X_d$.

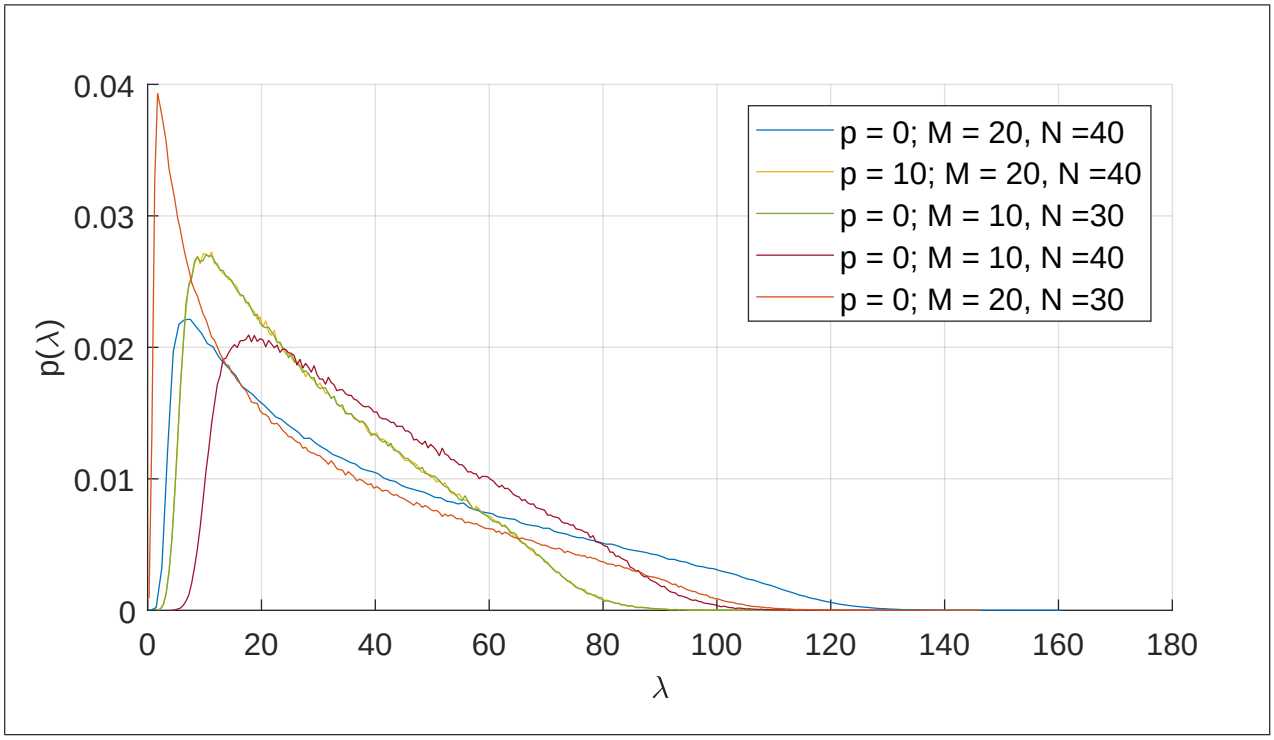

Figure 3: How should M and N change with p signal components? The overlap of yellow and green curves clearly shows both should decrease by p. (Simulation example with $p = 10$, $M = 20$, $N = 40$, $\sigma = 1$, SNR= $1e3$, and 1e5 repetitions.)

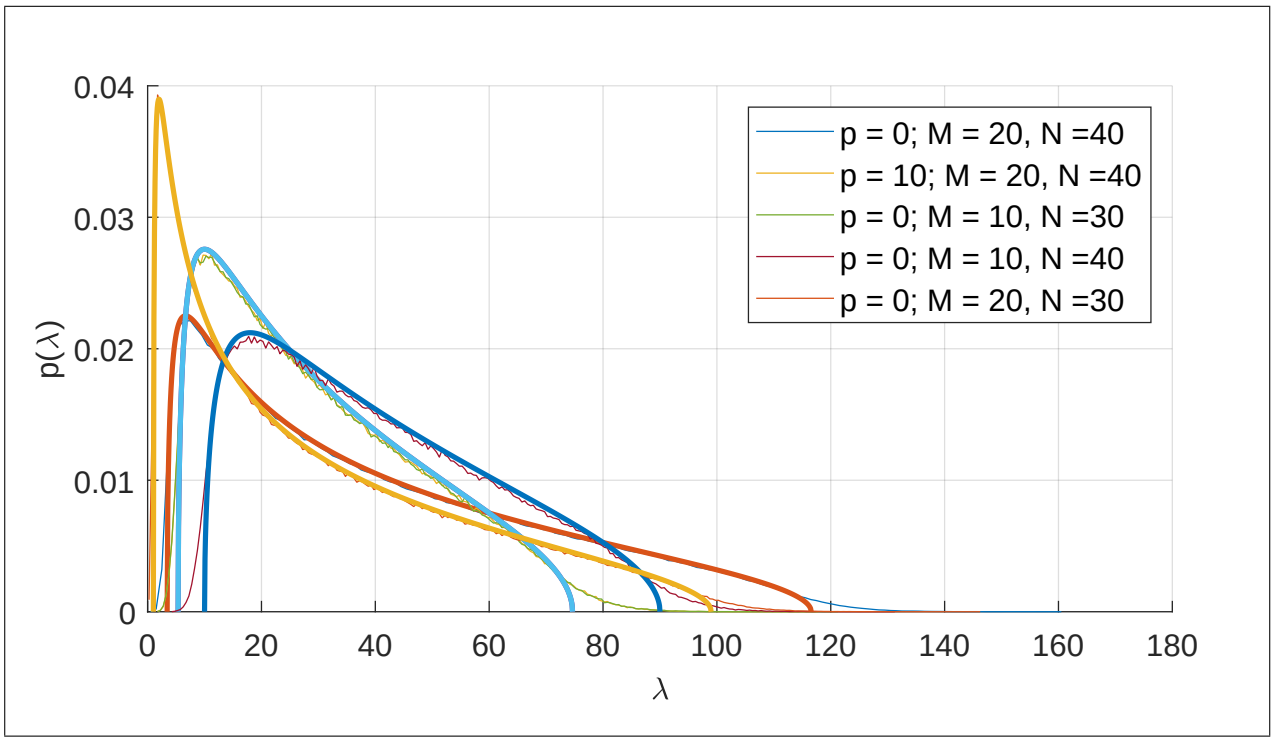

Figure 4: Same as previous but with added MP distribrutions.