The statistics of Principal Component Analysis

Olivier Catoni

CNRS, INRIA (projet CLASSIC)
Département de Mathématiques et Applications,
ENS, 45 rue d'Ulm, 75 230 Paris Cedex 05,
Olivier.Catoni@ens.fr

Toulouse, Université Paul Sabatier,
Tuesday, March 26, 2013

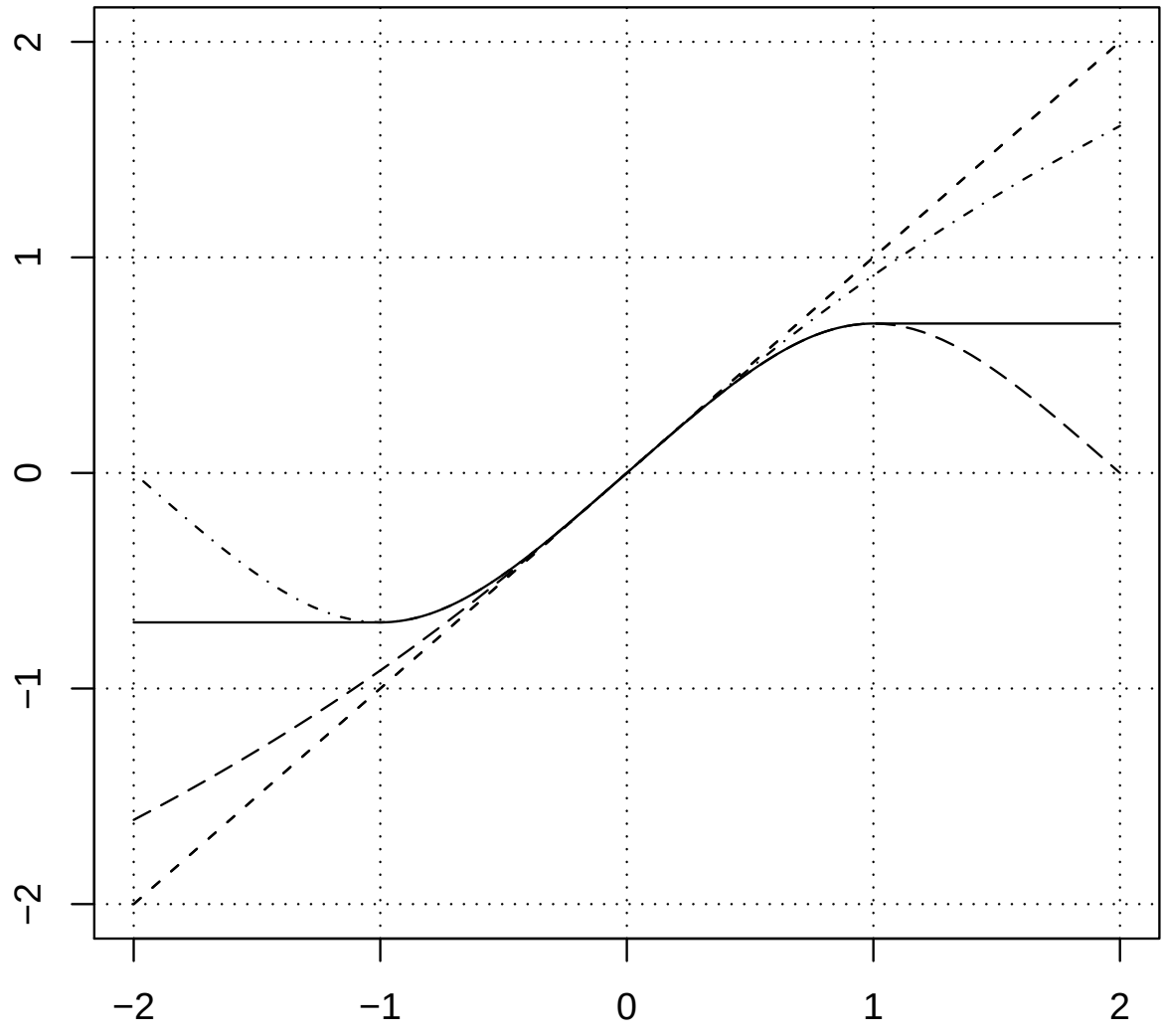
The statistics of PCA: joint work with Ilaria Giulini.

Principal component analysis in large dimension

Let us assume that $X_i \in \mathbb{R}^d$, where d may be large and $X_i \sim \mathbb{Q}$.
Question : Estimate $\int \langle x, \theta \rangle^2 \, \mathrm{d}\mathbb{Q}(x)$ for all $\theta \in \mathbb{R}^d$, or equivalently all $\theta \in S_d$, the sphere of $\mathbb{R}^d$.
We will give both dimension dependent and dimension independent PAC-Bayes bounds.

Let us introduce the influence function

$$\psi(z) = \begin{cases} \log(2), & z \geq 1, \\ -\log(1 - z + z^2/2), & 0 \leq z \leq 1, \\ -\psi(-z), & z \leq 0. \end{cases}$$



$z \mapsto \psi(z)$, compared with $z \mapsto z$
$z \mapsto \log(1+z+z^2/2)$, and $z \mapsto -\log(1-z+z^2/2)$

It is symmetric, non decreasing, bounded and satisfies

$$-\log\bigl(\min\{\log(2), 1 - z + z^2/2\}\bigr) \le \psi(z) \le \log\bigl(\min\{\log(2), 1 + z + z^2/2\}\bigr), \qquad z \in \mathbb{R}.$$

Dimension dependent bounds

Let $\overline{\mathbb{P}} = \frac{1}{n}\delta_{X_i}$ and

$$r_\lambda(\theta) = \lambda^{-1} \int \psi\big[\lambda\big[\langle\theta, x\rangle^2 - 1)\big]\,\mathrm{d}\overline{\mathbb{P}}(x),$$

where $\lambda > 0$ will be chosen later. Remark that

$$\int \lim_{\lambda\to 0} r_\lambda(\theta)\,\mathrm{d}\mathbb{P}^{\otimes n} = N(\theta) - 1.$$

We may assume that $G \stackrel{\text{def}}{=} \int xx^t\,\mathrm{d}\mathbb{P}(x)$ is invertible, since $\mathbb{P}\big(X \in \mathbf{Im}(G)\big) = 1$ and we can replace $\mathbb{R}^d$ with $\mathbf{Im}(G)$. Let

$$\widehat{\alpha}(\theta) = \sup\big\{\alpha \in \mathbb{R}_+ : r_\lambda(\alpha\theta) \le 0\big\}, \qquad \widehat{N}(\theta) = \widehat{\alpha}^{-2}.$$

$$\text{Let } s_4 = \left(\int \| G^{-1/2} x \|^4 \, \mathrm{d}\mathbb{P}(x) \right)^{1/4},$$

$$\kappa = \sup \Big\{ \int \langle \theta, x \rangle^4 \, \mathrm{d}\mathbb{P}(x), \theta \in \mathbb{R}^d, \int \langle \theta, x \rangle^2 \, \mathrm{d}\mathbb{P}(x) = 1 \Big\},$$

$$\lambda = \sqrt{\frac{2}{(\kappa - 1)n} \left[\log(\epsilon^{-1}) + \frac{(1 + 18c) s_4^2}{4\sqrt{\kappa}(1 + 4c)} \right]},$$

$$\eta = (\kappa - 1)\lambda = \sqrt{\frac{2(\kappa - 1)}{n} \left[\log(\epsilon^{-1}) + \frac{(1 + 18c) s_4^2}{4\sqrt{\kappa}(1 + 4c)} \right]}$$

$$\gamma = 2\sqrt{\frac{(1 + 4c) s_4^2 \sqrt{\kappa}}{n}},$$

$$\mu = \gamma + \eta,$$

$$\xi = \frac{\kappa \eta}{2(\kappa - 1)}.$$

Proposition

If $2\mu + \xi < 1$, *which is the case when*

$$n > \left(4\kappa^{1/4} s_4 \sqrt{1+4c} + 2\big(1+\kappa/(\kappa-1)\big) \right.$$
$$\left. \times \sqrt{2(\kappa-1)\left[\log(\epsilon^{-1}) + \frac{(1+18c)s_4^2}{4\sqrt{\kappa}(1+4c)}\right]} \right).$$

With probability at least $1-2\epsilon$, *for any* $\theta \in \mathbb{R}^d$,

$$\left| N(\theta)/\widehat{N}(\theta) - 1 \right| \leq \frac{\mu}{1-2\mu}.$$

Dimension free bounds

With probability $1-2\epsilon$, for any $\theta \in S_d$, and some estimator $\widehat{N}$ to be described in the proof,

$$\mathbb{1}\left(4\mu < 1\right)\left|N(\theta) - \widehat{N}(\theta)\right| \leq N(\theta)\frac{\mu}{1-4\mu},$$

where

$$\mu = \left(2\sqrt{\frac{(1+4c)s_4^2\sqrt{\kappa}}{nN(\theta)}} + \sqrt{\frac{2(\kappa-1)}{n}\left[\log\left(\frac{g+1}{\epsilon}\right) + \frac{(1+18c)s_4^2}{4(1+4c)N(\theta)\sqrt{\kappa}}\exp\left(\frac{\log(n)}{2g}\right)\right]}\right) \times \cosh\left(\frac{\log(n)}{2g}\right),$$

where $g \in \mathbb{N}$ is a grid parameter (you can take $g = \log(n)$ for instance) and this time $s_4 = \left(\int \|x\|^4 \mathrm{d}\mathbb{P}(x)\right)^{1/4}$.

Proof

Let $r_\lambda(\theta) = \int \psi(\langle\theta, x\rangle^2 - \lambda)\, \mathrm{d}\overline{\mathbb{P}}(x)$. Consider the Gaussian parameter perturbations $\pi_\theta = \mathscr{N}(\theta, \beta^{-1}\mathbb{I}_d)$, where $\mathbb{I}_d$ is the identity matrix of size $d \times d$. Let

$$\widehat{\alpha}(\theta) = \sup\{\alpha \in \mathbb{R}_+ : r_\lambda(\alpha\theta) \leq 0\}.$$

Proposition

Let $c = \frac{15}{\log(4)} \leq 11$.

$$\psi(\langle\theta, x\rangle^2 - \lambda) \leq \int \log\Bigg[1 + \langle\theta', x\rangle^2 - \lambda - \frac{\|x\|^2}{\beta} + \frac{1}{2}\left(\langle\theta', x\rangle^2 - \lambda - \frac{\|x\|^2}{\beta}\right)^2 + \frac{2c\|x\|^2}{\beta}\left(4\langle\theta', x\rangle^2 + \frac{5\|x\|^2}{\beta}\right)\Bigg] \mathrm{d}\pi_\theta(\theta').$$

Indeed,

$$\psi\Big(\int h\,\mathrm{d}\rho\Big) \leq \int \psi(h)\,\mathrm{d}\rho + \min\{\log(4), \mathbf{Var}(h\mathrm{d}\rho)\},$$

because $y \mapsto \psi(y) + \big(y - \int h\,\mathrm{d}\rho\big)^2$ is convex, since $\psi''(y) \geq -2$. As moreover

$$\langle\theta, x\rangle^2 - \lambda = \int \langle\theta', x\rangle^2\,\mathrm{d}\pi_\theta(\theta') - \lambda - \frac{\|x\|^2}{\beta},$$

we get

$$\psi(\langle\theta, x\rangle^2 - \lambda \leq \int \psi\Big(\langle\theta', x\rangle^2 - \frac{\|x\|^2}{\beta} - \lambda\Big)\,\mathrm{d}\pi_\theta(\theta') + \min\Big\{\log(4), \frac{4\|x\|^2\langle\theta, x\rangle^2}{\beta} + \frac{2\|x\|^4}{\beta^2}\Big\}$$

Lemma

If $W \sim \mathcal{N}(0, \sigma^2)$,

$$\min\{a, bm^2 + c\} \leq \mathbb{E}\Big(\min\{2a, 2b(m + W)^2 + 2b\sigma^2 + c\}\Big),$$

$$a, b, c \in \mathbb{R}_+, m \in \mathbb{R}.$$

The proof of this lemma is based on the inequalities $m^2 \leq 2(m + W)^2 + 2\,W^2$ and

$$\min\{a, y + z\} \leq \min\{a, y\} + \min\{a, z\}$$

$$\leq \min\{2a, y + z\}, \quad a, y, z \in \mathbb{R}_+.$$

Accordingly

$$\psi(\langle\theta,x\rangle^2-\lambda)\le\int\psi\bigg(\langle\theta',x\rangle^2-\frac{\|x\|^2}{\beta}-\lambda\bigg)\,\mathrm{d}\pi_\theta(\theta')$$
$$+\int\min\bigg\{4\log(2),\frac{8\|x\|^2\langle\theta',x\rangle^2}{\beta}+\frac{10\|x\|^4}{\beta^2}\bigg\}\,\mathrm{d}\pi_\theta(\theta').$$

We will now use

Lemma

For any $a, b, y, \in \mathbb{R}_+$ *and* $c = \frac{a}{b}\left[\exp(b) - 1\right]$,

$$\log(a) + \min\{b, y\} \leq \log(a + cy).$$

Applying this lemma to $a \leq 2$, $b = 4\log(2)$, and the corresponding $c = \frac{15}{\log(4)}$ ends the proof of the previous proposition.

PAC-Bayes bound

Proposition

For any measure $\nu \in \mathscr{M}_+^1(\Theta)$*, real number* $a > -1$ *and any measurable function* $f : \mathscr{X} \times \Theta \to [a, +\infty[$*, with probability at least* $1-\epsilon$*, for any* $\rho \in \mathscr{M}_+^1(\Theta)$ *such that* $\mathscr{K}(\rho, \nu) < \infty$*,*

$$\int \log\left[1+f(x,\theta)\right] \mathrm{d}\rho(\theta)\, \mathrm{d}\overline{\mathbb{P}}(x) \leq \int f(x,\theta)\, \mathrm{d}\rho(\theta)\, \mathrm{d}\mathbb{P}(x) + \frac{\mathscr{K}(\rho,\nu) - \log(\epsilon)}{n}.$$

The proof is based on the fact that

$$\int h \mathrm{d}\rho - \mathscr{K}(\rho,\nu) \leq \log\left(\int \exp(h) \mathrm{d}\nu\right),$$

for any upper-bounded measurable function $h : \Theta \to \mathbb{R}$.

We get with probability $1-\epsilon$, for any $\theta \in \mathbb{R}^d$,

$$\int \psi(\langle \theta, x\rangle^2 - \lambda)\, \mathrm{d}\overline{\mathbb{P}}(x) \leq \int \Bigg[\langle \theta, x\rangle^2 - \lambda + \frac{1}{2}\left(\left(\langle \theta, x\rangle^2 - \lambda\right)^2 + \frac{4}{\beta}\langle \theta, x\rangle^2 \|x\|^2 + \frac{2}{\beta^2}\|x\|^4 \right) + \frac{2c\|x\|^2}{\beta}\left(\frac{9\|x\|^2}{\beta} + 4\langle \theta, x\rangle^2 \right) \Bigg] \mathrm{d}\mathbb{P}(x) + \frac{\beta\|\theta\|^2}{2n} + \frac{\log(\epsilon^{-1})}{n}.$$

Using the Cauchy-Schwartz inequality will make the following quantities appear

$$
\begin{aligned}
s_4 &= \left(\int \|x\|^4 \, \mathrm{d}\mathbb{P}(x) \right)^{1/4}, \\
\kappa &= \sup \left\{ \int \langle \theta, x \rangle^4 \, \mathrm{d}\mathbb{P}(x), \theta \in \mathbb{R}^d \text{ s. t. } \int \langle \theta, x \rangle^2 \, \mathrm{d}\mathbb{P}(x) = 1 \right\}, \\
\xi &= \frac{\kappa \lambda}{2}, \\
\gamma &= \lambda(\kappa - 1) + \frac{2}{\beta}(1 + 4c) s_4^2 \sqrt{\kappa}, \\
\eta &= \frac{\lambda}{2}(\kappa - 1) + \frac{2}{\beta}(1 + 4c) s_4^2 \sqrt{\kappa} + \frac{(1 + 18c) s_4^4}{\beta^2 \lambda} - \frac{\log\left[\nu(\lambda, \beta)\epsilon\right]}{n\lambda}, \\
\delta &= \frac{\beta}{2n\lambda}.
\end{aligned}
$$

Proposition

With probability at least $1-2\epsilon$, *for any* $\theta \in \mathbb{R}^d$,

$$\frac{r_\lambda(\alpha\theta)}{\lambda} \leq \xi\left(\frac{N(\theta)}{\lambda}\alpha^2 - 1\right)^2 + (1+\gamma)\left(\frac{N(\theta)}{\lambda}\alpha^2 - 1\right) + \eta + \delta\|\theta\|^2\alpha^2,$$

$$\frac{r_\lambda(\alpha\theta)}{\lambda} \geq -\xi\left(\frac{\alpha^2 N(\theta)}{\lambda} - 1\right)^2 + \left(1-\gamma-\frac{\lambda\|\theta\|^2\delta}{N(\theta)}\right)\left(\frac{N(\theta)}{\lambda}\alpha^2 - 1\right) - \eta - \frac{\lambda\|\theta\|^2\delta}{N(\theta)}.$$

Let

$$\Phi_-(z) = z\left(1 - \frac{\eta + \frac{\delta\lambda}{z}}{1+\gamma-\eta-\frac{\delta\lambda}{z}}\right)\mathbb{1}\left(\xi - \gamma + \eta + \frac{\delta\lambda}{z} < 1\right),$$

$$\Phi_+(z) = z\left(1 + \frac{\eta + \frac{\delta\lambda}{z}}{1-\gamma-\eta-\frac{2\delta\lambda}{z}}\right)^{-1}\mathbb{1}\left(\xi + \gamma + \eta + \frac{2\delta\lambda}{z} < 1\right).$$

Replacing $N(\theta)$ with $\Phi_-\left(\lambda/\widehat{\alpha}^2\right)$ in the first inequality of the previous proposition and $\lambda/\widehat{\alpha}^2$ with $\Phi_+\left[N(\theta)\right]$ in the second inequality, we can prove that

$$\Phi_-\left(\frac{\lambda}{\widehat{\alpha}^2}\right) \leq N(\theta),$$

$$\Phi_+\left[N(\theta)\right] \leq \frac{\lambda}{\widehat{\alpha}^2}.$$

Proposition

With probability at least $1-2\epsilon$, *for any* $\theta \in \mathbb{R}^d$,

$$B_- = \sup_{\lambda,\beta} \Phi_-\left(\frac{\lambda}{\widehat{\alpha}^2}\right) \leq N(\theta) \leq \inf_{\lambda,\beta} \Phi_+\left(\frac{\lambda}{\widehat{\alpha}^2}\right) = B_+.$$

Let us put $\widehat{N}(\theta) = \dfrac{B_- + B_+}{2}$, we get with probability at least $1-2\epsilon$, for any $\lambda, \beta \in \mathbb{R}_+$,

$$N(\theta) - \widehat{N}(\theta) \leq \frac{N - B_-}{2} = \frac{N - \Phi_-\left(\lambda/\widehat{\alpha^2}\right)}{2} \leq \frac{N(\theta) - \Phi_- \circ \Phi_+\left[N(\theta)\right]}{2},$$

$$\widehat{N}(\theta) - N(\theta) \leq \frac{\Phi_+^{-1}\left(\lambda/\widehat{\alpha}^2\right) - N(\theta)}{2} \leq \frac{\Phi_+^{-1} \circ \Phi_-^{-1}\left[N(\theta)\right] - N(\theta)}{2}.$$

Putting $r(z) = \dfrac{\eta + \frac{\delta\lambda}{z}}{1-\gamma-\eta-\frac{2\delta\lambda}{z}}$, and $c(z) = \eta + \dfrac{\delta\lambda}{z}$, we can prove that

$$\Phi_-(z) \geq z\big[1-r(z)\big]\mathbb{1}\big[4c(z)<1\big],$$
$$\Phi_+(z) \geq z\big[1+r(z)\big]^{-1}\mathbb{1}\big[4c(z)<1\big]$$
$$\Phi_-^{-1}(z)\mathbb{1}\big[4c(z)<1\big] \leq z\big[1-r(z)\big]^{-1},$$
$$\Phi_+^{-1}(z)\mathbb{1}\big[4c(z)<1\big] \leq z\big[1+r(z)\big].$$

From this we can deduce

Proposition

With probablity at least $1-2\epsilon$, *for any* $\theta \in \mathbb{R}^d$,

$$\mathbb{1}\big[4c(N(\theta)) < 1\big]\Big|N(\theta) - \widehat{N}(\theta)\Big| \leq N(\theta)\frac{c\big(N(\theta)\big)}{1-4c\big(N(\theta)\big)}.$$

The announced result is then obtained by optimizing the value of $c\big(N(\theta)\big)$ by appropriate choices of λ and β.