

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

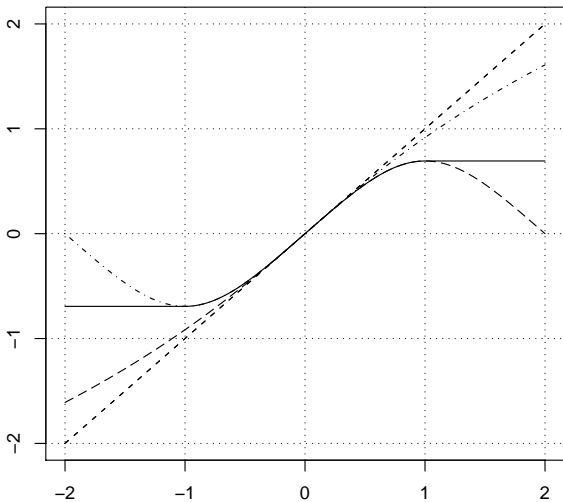
Let us assume that $X_i \in \mathbb{R}^d$, where d may be large and $X_i \sim \mathbb{Q}$.

Question : Estimate $\int \langle x, \theta \rangle^2 d\mathbb{Q}(x)$ 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

$$\begin{aligned} -\log(\min\{\log(2), 1 - z + z^2/2\}) &\leq \psi(z) \\ &\leq \log(\min\{\log(2), 1 + z + z^2/2\}), \quad z \in \mathbb{R}. \end{aligned}$$

Dimension dependent bounds

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

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

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

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

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

$$\hat{\alpha}(\theta) = \sup\{\alpha \in \mathbb{R}_+ : r_\lambda(\alpha\theta) \leq 0\}, \quad \hat{N}(\theta) = \hat{\alpha}^{-2}.$$

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

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

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

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

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

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

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

Proposition

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

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

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

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

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,

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

where

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

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

Proof

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

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

Proposition

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

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

Indeed,

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

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

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

we get

$$\begin{aligned} \psi(\langle \theta, x \rangle^2 - \lambda) &\leq \int \psi\left(\langle \theta', x \rangle^2 - \frac{\|x\|^2}{\beta} - \lambda\right) \, d\pi_\theta(\theta') \\ &\quad + \min\left\{\log(4), \frac{4\|x\|^2 \langle \theta, x \rangle^2}{\beta} + \frac{2\|x\|^4}{\beta^2}\right\} \end{aligned}$$

Lemma

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

$$\min\{a, bm^2 + c\} \leq \mathbb{E}\left(\min\{2a, 2b(m + W)^2 + 2b\sigma^2 + c\}\right),$$
$$a, b, c \in \mathbb{R}_+, m \in \mathbb{R}.$$

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

$$\begin{aligned} \psi(\langle \theta, x \rangle^2 - \lambda) &\leq \int \psi\left(\langle \theta', x \rangle^2 - \frac{\|x\|^2}{\beta} - \lambda\right) d\pi_{\theta}(\theta') \\ &+ \int \min\left\{4\log(2), \frac{8\|x\|^2\langle \theta', x \rangle^2}{\beta} + \frac{10\|x\|^4}{\beta^2}\right\} d\pi_{\theta}(\theta'). \end{aligned}$$

We will now use

Lemma

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

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

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

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

The proof is based on the fact that

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

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

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

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

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

$$s_4 = \left(\int \|x\|^4 d\mathbb{P}(x) \right)^{1/4},$$

$$\kappa = \sup \left\{ \int \langle \theta, x \rangle^4 d\mathbb{P}(x), \theta \in \mathbb{R}^d \text{ s. t. } \int \langle \theta, x \rangle^2 d\mathbb{P}(x) = 1 \right\},$$

$$\xi = \frac{\kappa\lambda}{2},$$

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

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

$$\delta = \frac{\beta}{2n\lambda}.$$

Proposition

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

$$\begin{aligned}\frac{r_\lambda(\alpha\theta)}{\lambda} &\leq \xi \left(\frac{N(\theta)}{\lambda} \alpha^2 - 1 \right)^2 + (1 + \gamma) \left(\frac{N(\theta)}{\lambda} \alpha^2 - 1 \right) + \eta + \delta \|\theta\|^2 \alpha^2, \\ \frac{r_\lambda(\alpha\theta)}{\lambda} &\geq -\xi \left(\frac{\alpha^2 N(\theta)}{\lambda} - 1 \right)^2 + \left(1 - \gamma - \frac{\lambda \|\theta\|^2 \delta}{N(\theta)} \right) \left(\frac{N(\theta)}{\lambda} \alpha^2 - 1 \right) \\ &\quad - \eta - \frac{\lambda \|\theta\|^2 \delta}{N(\theta)}.\end{aligned}$$

Let

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

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

$$\begin{aligned}\Phi_-\left(\frac{\lambda}{\widehat{\alpha}^2}\right) &\leq N(\theta), \\ \Phi_+[N(\theta)] &\leq \frac{\lambda}{\widehat{\alpha}^2}.\end{aligned}$$

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

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

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

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

$$\Phi_-(z) \geq z[1 - r(z)] \mathbf{1}[4c(z) < 1],$$

$$\Phi_+(z) \geq z[1 + r(z)]^{-1} \mathbf{1}[4c(z) < 1]$$

$$\Phi_-^{-1}(z) \mathbf{1}[4c(z) < 1] \leq z[1 - r(z)]^{-1},$$

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

From this we can deduce

Proposition

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

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

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