Principal Component Analysis for a Spiked Covariance Model with Largest Eigenvalues of the Same Asymptotic Order of Magnitude

Addy M. Bolívar Cimé

Centro de Investigación en Matemáticas A.C.

May 1, 2010

Contenido

Contenido

Contenido

An important problem in multivariate statistical analysis is the estimation of the population covariance matrix. In principal component analysis (PCA) one often fails to estimate the population eigenvalues and eigenvectors, since the sample covariance matrix is not a good approximation to the population covariance matrix when the data dimension is larger than the sample size.

Spiked Covariance Model

As pointed out in Johnstone (2001), one often observes one or a small number of large sample eigenvalues well separated from the rest. In this case, of special interest is the so-called *spiked covariance model*.

More specifically, suppose we have a $d \times n$ data matrix $X = [x_1, x_2, \ldots, x_n]$ with $d \gg n$, in the sense that $\frac{d}{n} \to \infty$, where $x_j = (x_{1j}, \ldots, x_{dj})^\top$, $j = 1, 2, \ldots, n$. Assume the columns of X are independent and identically distributed random vectors from a multivariate Gaussian distribution with mean zero and unknown covariance matrix Σ. The *spiked covariance model* considers a covariance matrix of the type

$$\Sigma = \operatorname{diag}(\tau_1, \tau_2, \ldots, \tau_p, 1, \ldots, 1), \tag{1}$$

with $\tau_1 \geq \tau_2 \geq \cdots \geq \tau_p > 1$, for some $1 \leq p < d$.

Some results

Ahn, Marron, Muller and Chi (2007) showed that for $p = 1$ and $\Sigma = \mathrm{diag}(d^{\alpha}, 1, \ldots, 1)$, with $\alpha > 1$, it is possible to estimate very well the first population eigenvalue and eigenvector using PCA when d and n are sufficiently large and $d \gg n$. For the case $p \geq 2$, Jung and Marron (2009) showed the consistency of the first p largest sample eigenvalues under the assumption that each of the p largest population eigenvalues, $\tau_1 \geq \tau_2 \geq \cdots \geq \tau_p$, has a different asymptotic order of magnitude when dimension increases, that is

$$\frac{\tau_i}{d^{\alpha_i}} \to c_i, \qquad \text{as } d \to \infty,$$

where $\alpha_1 > \alpha_2 > \cdots > \alpha_p > 1$ and $c_i > 0$, $i = 1, 2, \ldots, p$.

They showed that if $\widehat{\tau}_1 \geq \widehat{\tau}_2 \geq \cdots \geq \widehat{\tau}_p$ are the p largest sample eigenvalues

$$\frac{\widehat{\tau}_i}{\tau_i} \xrightarrow{w} \frac{\chi^2_n}{n}, \quad \text{as } d \to \infty, \tag{2}$$

for $i = 1, 2, \ldots, p$, where χ^2_n is a Chi-square random variable with n degrees of freedom. Therefore, since $\chi^2_n/n \xrightarrow{w} 1$ as $n \to \infty$ the first p sample eigenvalues are consistent. They also show that in this case the corresponding sample eigenvectors are consistent.

Purpose of this work

To study the asymptotic behaviour of the sample eigenvalues and eigenvectors, when d, n are sufficiently large and $d \gg n$, in the case of the spiked covariance model where the p largest population eigenvalues $\tau_1, \ldots, \tau_p$, $p > 1$, have the same asymptotic order of magnitude when dimension increases, i.e. $\alpha_1 = \alpha_2 = \cdots = \alpha_p = \alpha > 1$ and $c_1 = c_2 = \cdots = c_p = c > 0$.

Under this last assumption, we first prove a multivariate extension of (2), namely

$$(\frac{\widehat{\tau}_1}{\tau_1}, \frac{\widehat{\tau}_2}{\tau_2}, \ldots, \frac{\widehat{\tau}_p}{\tau_p})^\top \xrightarrow{w} (cn)^{-1}(\ell_1, \ell_2, \ldots, \ell_p)^\top \quad \text{as } d \to \infty$$

where $\ell_1 \geq \ell_2 \geq \cdots \geq \ell_p$ are the nonzero eigenvalues of a Wishart random matrix with p degrees of freedom and $n \times n$ covariance matrix $c\mathrm{I}_n$.

Wishart Distribution

Lemma 2.1

If $\ell_1 \geq \ell_2 \geq \cdots \geq \ell_p$ are the nonzero eigenvalues of $U \sim \mathcal{W}_n(p, cI_n)$ with $n > p$ and $c > 0$, then the joint density of $(\ell_1, \ell_2, \ldots, \ell_p)^\top$ is given by

$$f_{\ell_1,\ell_2,\ldots,\ell_p}(\ell_1, \ell_2, \ldots, \ell_p) = \Delta \sum_{\alpha \in S_p} \operatorname{sign}(\alpha) \exp\left(-\frac{1}{2c}\sum_{k=1}^{p} \ell_k\right) * \prod_{k=1}^{p} \ell_{p+1-k}^{\alpha_k + (n-p-1)/2 - 1}, \tag{3}$$

with $\ell_1 > \ell_2 > \cdots > \ell_p$, where

$$\Delta = \frac{\pi^{p^2/2}}{(2c)^{pn/2}\Gamma_p(\frac{p}{2})\Gamma_p(\frac{n}{2})}. \tag{4}$$

Proposition 2.1

If $\ell_1 \geq \ell_2 \geq \cdots \geq \ell_p$ are the nonzero eigenvalues of $U \sim \mathcal{W}_n(p, cI_n)$ with $n > p$ and $c > 0$, then the characteristic function of $(\ell_1, \ell_2, \ldots, \ell_p)^\top$ is given by

$$\varphi_{\ell_1,\ell_2,\ldots,\ell_p}(t_1, t_2, \ldots, t_p) =$$

$$\Delta \left(\frac{2c}{p}\right)^{pn/2} \widehat{G}(\sum_{j=1}^{p} t_j; \frac{pn}{2}, \frac{2c}{p}) \sum_{k_1=0}^{\infty} \cdots \sum_{k_{p-1}=0}^{\infty} \Gamma(\frac{pn}{2} + \sum_{j=1}^{p-1} k_j)$$

$$* \, C_{n,p}(k_1, \ldots, k_{p-1}) \prod_{r=1}^{p-1} \left(\frac{r}{p}\right)^{k_r} \frac{\widehat{G}(\sum_{j=1}^{p} t_j; k_r, \frac{2c}{p})}{\widehat{G}(\sum_{j=p+1-r}^{p} t_j; k_r, \frac{2c}{r})}, \tag{5}$$

where $\widehat{G}(t; a, b)$ is the characteristic function of the gamma distribution Gamma(a, b) and

$$C_{n,p}(k_1, \ldots, k_{p-1}) = \sum_{\alpha \in S_p} \operatorname{sign}(\alpha) \prod_{r=1}^{p-1} \frac{\Gamma(\sum_{j=1}^{r} \alpha_j + \frac{r(n-p-1)}{2} + \sum_{j=1}^{r-1} k_j)}{\Gamma(\sum_{j=1}^{r} \alpha_j + \frac{r(n-p-1)}{2} + \sum_{j=1}^{r} k_j + 1)}. \tag{6}$$

Theorem 3.1

Assume X as above with unknown covariance matrix given by the spiked covariance model (1) with $p < n < d$ and where $\tau_1 \geq \tau_2 \geq \cdots \geq \tau_p > 1$ have the same asymptotic order of magnitude, that is $\tau_i = \tau_i(d)$, $d^{-\alpha}\tau_i(d) \to c$ as $d \to \infty$ for some $c > 0$ and $\alpha > 1$, $i = 1, 2, \ldots, p$. Let $\widehat{\tau}_1 \geq \widehat{\tau}_2 \geq \cdots \geq \widehat{\tau}_p$ be the p largest sample eigenvalues of the sample covariance matrix $S = n^{-1}XX^\top$. Then

$$\left(\frac{\widehat{\tau}_1}{\tau_1}, \frac{\widehat{\tau}_2}{\tau_2}, \ldots, \frac{\widehat{\tau}_p}{\tau_p}\right)^\top \xrightarrow{w} (cn)^{-1}(\ell_1, \ell_2, \ldots, \ell_p)^\top,$$

when $d \to \infty$, where $\ell_1 \geq \ell_2 \geq \cdots \geq \ell_p$ are the nonzero eigenvalues of a Wishart random matrix with distribution $\mathcal{W}_n(p, cI_n)$.

Proposition 3.1

Let $c > 0$ and $\ell_1 \geq \ell_2 \geq \cdots \geq \ell_p$ be the nonzero eigenvalues of $U \sim \mathcal{W}_n(p, cI_n)$ with $n > p$. Then

$$(cn)^{-1}(\ell_1, \ell_2, \ldots, \ell_p)^\top \xrightarrow{w} (1, 1, \ldots, 1)^\top, \qquad \text{as } n \to \infty.$$

Theorem 3.2

Under the assumptions of Theorem 3.1, we have that for all $\epsilon > 0$

$$\lim_{n\to\infty} \lim_{d\to\infty} P\left(\| \left(\frac{\widehat{\tau}_1}{\tau_1}, \frac{\widehat{\tau}_2}{\tau_2}, \ldots, \frac{\widehat{\tau}_p}{\tau_p}\right)^\top - (1, 1, \ldots, 1)^\top \| > \epsilon \right) = 0. \qquad (7)$$

Theorem 3.3

Under the same conditions of Theorem 3.1 we have

$$d^{-\alpha} n(\widehat{\tau}_1, \widehat{\tau}_2, \ldots, \widehat{\tau}_n)^\top \xrightarrow{w} (\ell_1, \ell_2, \ldots, \ell_p, 0, \ldots, 0)^\top$$

as $d \to \infty$, where $\ell_1 \geq \ell_2 \geq \cdots \geq \ell_p$ are the nonzero eigenvalues of a Wishart random matrix with distribution $\mathcal{W}_n(p, cI_n)$.

We define

$$\text{Angle}(v_j, \text{span}\{u_i : i \in J\}) = \arccos\left(\frac{v_j^\top [\text{Proj}_{\text{span}\{u_i : i \in J\}} v_j]}{\| v_j \| \| \text{Proj}_{\text{span}\{u_i : i \in J\}} v_j \|}\right)$$
$$= \arccos\left(\frac{v_j^\top (\sum_{i \in J} (u_i^\top v_j) u_i)}{\| v_j \| \| \sum_{i \in J} (u_i^\top v_j) u_i \|}\right).$$

Following the definition given in Jung and Marron (2009), we say that the sample eigenvector v_j, with $j \in J$, is *subspace consistent* if

$$\text{Angle}(v_j, \text{span}\{u_i : i \in J\}) \xrightarrow{P} 0.$$

Theorem 3.4

Under the same conditions of Theorem 3.1, let $v_1, v_2, \ldots, v_p$ *be the sample eigenvectors corresponding to the first p sample eigenvalues* $\widehat{\tau}_1 \geq \widehat{\tau}_2 \geq \cdots \geq \widehat{\tau}_p$. *Then if* $\epsilon > 0$

$$\lim_{n\to\infty} \lim_{d\to\infty} P(\mathrm{Angle}(v_i, \mathrm{span}\{e_j : j = 1, 2, \ldots p\}) > \epsilon) = 0, \qquad (8)$$

for $i = 1, 2, \ldots, p$.