Robust estimation of principal components from depth-based multivariate rank covariance matrix

Subho Majumdar
Snigdhansu Chatterjee

University of Minnesota, School of Statistics

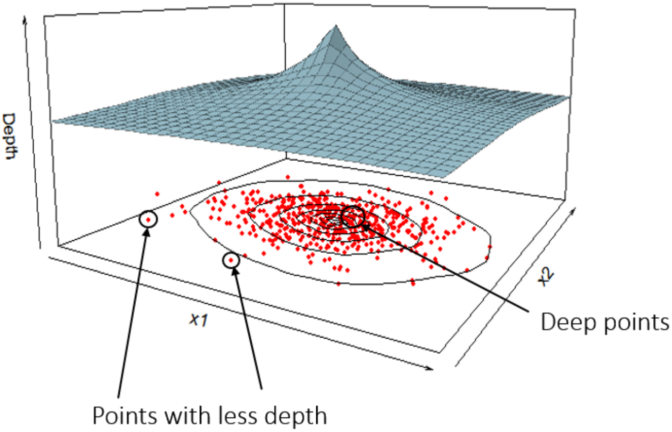
July 9, 2015

Table of contents

Summary

- Introduction: what is data depth?
- Multivariate ranks based on data depth
- The Depth Covariance Matrix (DCM): overview of results
- Performance: simulations and real data analysis

Example: 500 points from $\mathcal{N}_2((0,0)^T, \text{diag}(2,1))$



A scalar measure of how much inside a point is with respect to a data cloud

Formal definition

For any multivariate distribution $F = F_{\mathbf{X}}$, the depth of a point $\mathbf{x} \in \mathbb{R}^p$, say $D(\mathbf{x}, F_{\mathbf{X}})$ is any real-valued function that provides a 'center outward ordering' of $\mathbf{x}$ with respect to F (Zuo and Serfling, 2000).

Desirable properties (Liu, 1990)

(P1) *Affine invariance*: $D(A\mathbf{x} + \mathbf{b}, F_{A\mathbf{X}+\mathbf{b}}) = D(\mathbf{x}, F_{\mathbf{X}})$

(P2) *Maximality at center*: $D(\theta, F_{\mathbf{X}}) = \sup_{\mathbf{x} \in \mathbb{R}^p} D(\mathbf{x}, F_{\mathbf{X}})$ for $F_{\mathbf{X}}$ with center of symmetry θ, the *deepest point* of $F_{\mathbf{X}}$.

(P3) *Monotonicity w.r.t. deepest point*: $D(\mathbf{x}; F_{\mathbf{X}}) \leq D(\theta + a(\mathbf{x} - \theta), F_{\mathbf{X}})$

(P4) *Vanishing at infinity*: $D(\mathbf{x}; F_{\mathbf{X}}) \to \mathbf{0}$ as $\|\mathbf{x}\| \to \infty$.

- **Halfspace depth** (HD) (Tukey, 1975) is the minimum probability of all halfspaces containing a point.

$$HD(\mathbf{x}, F) = \inf_{\mathbf{u} \in \mathbb{R}^p; \mathbf{u} \neq \mathbf{0}} P(\mathbf{u}^T \mathbf{X} \geq \mathbf{u}^T \mathbf{x})$$

- **Projection depth** (PD) (Zuo, 2003) is based on an outlyingness function:

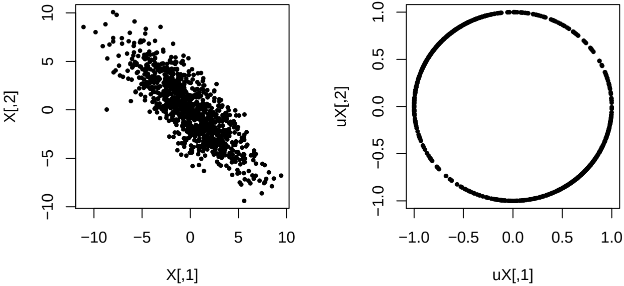
$$O(\mathbf{x}, F) = \sup_{\|\mathbf{u}\|=1} \frac{|\mathbf{u}^T \mathbf{x} - m(\mathbf{u}^T \mathbf{X})|}{s(\mathbf{u}^T \mathbf{X})}; \quad PD(\mathbf{x}, F) = \frac{1}{1 + O(\mathbf{x}, F)}$$

Robustness

- **Classification**
- Depth-weighted means and covariance matrices
- What we're going to do:
  PCA based on covariance matrix of depth-based multivariate rank vectors

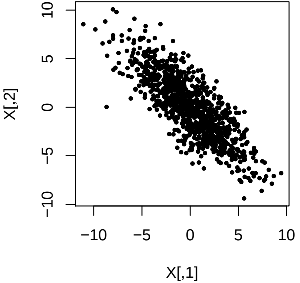
$$\mathbf{S}(\mathbf{x}) = \begin{cases} \mathbf{x}\|\mathbf{x}\|^{-1} & \text{if } \mathbf{x} \neq \mathbf{0} \\ \mathbf{0} & \text{if } \mathbf{x} = \mathbf{0} \end{cases}$$

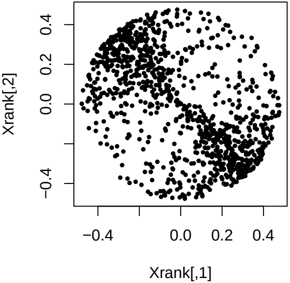
- Say $\mathbf{x}$ follows an elliptic distribution with mean μ, covariance matrix Σ.
- Sign covariance matrix (SCM): $\Sigma_S(\mathbf{X}) = E\mathbf{S}(\mathbf{X}-\mu)\mathbf{S}(\mathbf{X}-\mu)^T$
- SCM has same eigenvectors as Σ. PCA using SCM is robust, but not efficient.



Spatial ranks

- Fix a depth function $D(\mathbf{x}, F) = D_\mathbf{X}(\mathbf{x})$. Define $\tilde{D}_\mathbf{X}(\mathbf{x}) = \sup_{\mathbf{x} \in \mathbb{R}^p} D_\mathbf{X}(\mathbf{x}) - D_\mathbf{X}(\mathbf{x})$
- Transform the original observation: $\tilde{\mathbf{x}} = \tilde{D}_\mathbf{X}(\mathbf{x})\mathbf{S}(\mathbf{x} - \mu)$. This is the *Spatial Rank* of $\mathbf{x}$.
- Depth Covariance Matrix (DCM) = $Cov(\tilde{\mathbf{X}})$. Has more information than spatial signs, so more efficient.





Theorem (1)

Let the random variable $\mathbf{X} \in \mathbb{R}^p$ follow an elliptical distribution with center $\boldsymbol{\mu}$ and covariance matrix $\Sigma = \Gamma\Lambda\Gamma^T$, its spectral decomposition. Then, given a depth function $D_{\mathbf{X}}(.)$ the covariance matrix of the transformed random variable $\tilde{\mathbf{X}}$ is

$$Cov(\tilde{\mathbf{X}}) = \Gamma\Lambda_{D,S}\Gamma^T, \quad \textit{with} \quad \Lambda_{D,S} = E\left[(\tilde{D}_{\mathbf{Z}}(\mathbf{z}))^2 \frac{\Lambda^{1/2}\mathbf{z}\mathbf{z}^T\Lambda^{1/2}}{\mathbf{z}^T\Lambda\mathbf{z}}\right] \tag{1}$$

where $\mathbf{z} = (z_1, ..., z_p)^T \sim N(\mathbf{0}, I_p)$ and $\Lambda_{D,S}$ a diagonal matrix with diagonal entries

$$\lambda_{D,S,i} = E_{\mathbf{Z}}\left[\frac{(\tilde{D}_{\mathbf{Z}}(\mathbf{z}))^2\lambda_i z_i^2}{\sum_{j=1}^p \lambda_j z_j^2}\right]$$

Other results

- Asymptotic distribution of sample DCM, form of its asymptotic variance
- Asymptotic joint distribution of eigenvectors and eigenvalues of sample DCM
- Form and shape of influence function: a measure of robustness
- Asymptotic efficiency relative to sample covariance matrix

- 6 elliptical distributions: p-variate normal and t- distributions with df = 5, 6, 10, 15, 25.
- All distributions centered at $\mathbf{0}_p$, and have covariance matrix $\Sigma = \text{diag}(p, p-1, ...1)$.
- 3 choices of p: 2, 3 and 4.
- 10000 samples each for sample sizes $n = 20, 50, 100, 300, 500$
- For estimates $\hat{\gamma}_1$ of the first eigenvector $\hat{\gamma}_1$, prediction error is measured by the average smallest angle between the two lines, i.e. **Mean Squared Prediction Angle**:

$$MSPA(\hat{\gamma}_1) = \frac{1}{10000} \sum_{m=1}^{10000} \left(\cos^{-1} \left| \gamma_1^T \hat{\gamma}_1^{(m)} \right| \right)^2$$

Finite sample efficiency of some eigenvector estimate $\hat{\gamma}_1^E$ relative to that obtained from the sample covariance matrix, say $\hat{\gamma}_1^{Cov}$ is:

$$FSE(\hat{\gamma}_1^E, \hat{\gamma}_1^{Cov}) = \frac{MSPA(\hat{\gamma}_1^{Cov})}{MSPA(\hat{\gamma}_1^E)}$$

Table of FSE for $p = 2$

F = Bivariate t_5	SCM	HSD-CM	MhD-CM	PD-CM
n=20	0.80	0.95	0.95	0.89
n=50	0.86	1.25	1.10	1.21
n=100	1.02	1.58	1.20	1.54
n=300	1.24	1.81	1.36	1.82
n=500	1.25	1.80	1.33	1.84
F = Bivariate t_6	SCM	HSD-CM	MhD-CM	PD-CM
n=20	0.77	0.92	0.92	0.86
n=50	0.76	1.11	1.00	1.08
n=100	0.78	1.27	1.06	1.33
n=300	0.88	1.29	1.09	1.35
n=500	0.93	1.37	1.13	1.40
F = Bivariate t_{10}	SCM	HSD-CM	MhD-CM	PD-CM
n=20	0.70	0.83	0.84	0.77
n=50	0.58	0.90	0.84	0.86
n=100	0.57	0.92	0.87	0.97
n=300	0.62	0.93	0.85	0.99
n=500	0.62	0.93	0.86	1.00

Table of FSE for $p = 2$

F = Bivariate t_{15}	SCM	HSD-CM	MhD-CM	PD-CM
n=20	0.63	0.76	0.78	0.72
n=50	0.52	0.79	0.75	0.80
n=100	0.51	0.83	0.77	0.88
n=300	0.55	0.84	0.79	0.91
n=500	0.56	0.85	0.80	0.93
F = Bivariate t_{25}	SCM	HSD-CM	MhD-CM	PD-CM
n=20	0.63	0.77	0.79	0.74
n=50	0.49	0.73	0.71	0.76
n=100	0.45	0.73	0.69	0.81
n=300	0.51	0.78	0.75	0.87
n=500	0.53	0.79	0.75	0.87
F = BVN	SCM	HSD-CM	MhD-CM	PD-CM
n=20	0.56	0.69	0.71	0.67
n=50	0.42	0.66	0.66	0.70
n=100	0.42	0.69	0.66	0.77
n=300	0.47	0.71	0.69	0.82
n=500	0.48	0.73	0.71	0.83

- Features extracted from images of 213 buses: 18 variables
- Methods compared:

  Classical PCA (CPCA)
  SCM PCA (SPCA)
  ROBPCA (Hubert et al., 2005)
  PCA based on MCD (MPCA)
  PCA based on projection-DCM (DPCA)

q	Method of PCA				
	CPCA	SPCA	ROBPCA	MPCA	DPCA
1	0.188	0.549	0.410	0.514	**0.662**
2	0.084	0.272	0.214	0.337	**0.359**
3	0.044	0.182	0.121	0.227	**0.237**
4	0.026	0.135	0.083	0.154	**0.173**
5	0.018	0.099	0.054	0.098	**0.115**
6	0.012	0.069	0.036	0.070	**0.084**

Table : Unexplained proportions of variability by PCA models with q components for bus data

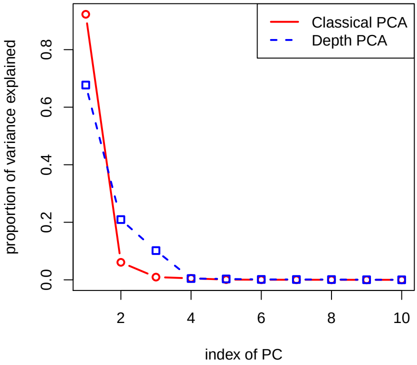
- Proportions of variability that are left unexplained after the top q $(= 1, ..., 6)$ components are taken into account,
- First PC of CPCA seems to explain a lot of variability as classical variances are inflated due to outliers in the direction of the first principal axis. Robust methods do not suffer from this.

Quantile	Method of PCA				
	CPCA	SPCA	ROBPCA	MPCA	DPCA
10%	1.9	1.2	1.2	1.0	**1.2**
20%	2.3	1.6	1.6	1.3	**1.6**
30%	2.8	1.8	1.8	1.7	**1.9**
40%	3.2	2.2	2.1	2.1	**2.3**
50%	3.7	2.6	2.5	3.1	**2.6**
60%	4.4	3.1	3.0	5.9	**3.2**
70%	5.4	3.8	3.9	25.1	**3.9**
80%	6.5	5.2	4.8	86.1	**4.8**
90%	8.2	9.0	10.9	298.2	**6.9**
Max	24	1037	1055	1037	**980**

Table : Quantiles to squared distance from 3-principal component hyperplanes for bus data

- Quantiles of the squared orthogonal distance for a sample point from the hyperplane formed by top 3 PCs,
- For DPCA, more than 90% of points have a smaller orthogonal distance than CPCA

- 226 variables and 39 observations. Each observation is a gasoline sample with a certain octane number, and have their NIR absorbance spectra measured in 2 nm intervals between 1100 - 1550 nm.
- 6 outliers: compounds 25, 26 and 36-39, which contain alcohol.



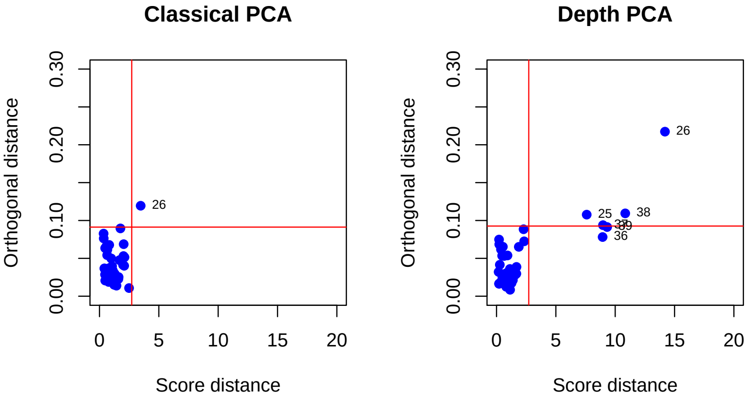


Figure : Distance plots for two types of PCA on octane data

Extensions: Robust kernel PCA

- 20 points from each person. Noise added to one image from each person.
- Columns due to kernel CPCA, SPCA and DPCA, respectively. Rows due to top 2, 4, 6, 8 or 10 PCs considered.

Extensions: future work

- Explore properties of a depth-weighted M-estimator of scale matrix:

$$\Sigma_{Dw} = E\left[\frac{(\tilde{D}_{\mathbf{X}}(\mathbf{x}))^2 (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T}{(\mathbf{x} - \boldsymbol{\mu})^T \Sigma_{Dw}^{-1} (\mathbf{x} - \boldsymbol{\mu})} \right]$$

- Leverage the idea of depth based ranks: robust non-parametric testing
- Extending to high-dimensional and functional data

M. Hubert, P. J. Rousseeuw, and K. V. Branden. ROBPCA: A New Approach to Robust Principal Component Analysis. *Technometrics*, 47-1:64–79, 2005.

R.Y. Liu. On a notion of data depth based on random simplices. *Ann. of Statist.*, 18:405–414, 1990.

N. Locantore, J.S. Marron, D.G. Simpson, N. Tripoli, J.T. Zhang, and K.L. Cohen. Robust principal components of functional data. *TEST*, 8:1–73, 1999.

J.W. Tukey. Mathematics and picturing data. In R.D. James, editor, *Proceedings of the International Congress on Mathematics*, volume 2, pages 523–531, 1975.

Y. Zuo. Projection-based depth functions and associated medians. *Ann. Statist.*, 31:1460–1490, 2003.

Y. Zuo and R. Serfling. General notions of statistical depth functions. *Ann. Statist.*, 28-2:461–482, 2000.

THANK YOU!