An automatic robust Bayesian approach to principal component regression – supplementary material

Philippe Gagnon[a], Mylène Bédard[b] and Alain Desgagné[c]

[a]Department of Statistics, University of Oxford, Oxford, United Kingdom; [b]Département de mathématiques et de statistique, Université de Montréal, Montréal, QC, Canada; [c]Département de mathématiques, Université du Québec à Montréal, Montréal, QC, Canada

ARTICLE HISTORY
Compiled December 16, 2019

We first present the mathematical justification of the approximate robust principal component analysis (PCA) in Section 1. The validity of our prior structure is next addressed in Section 2. Propositions 3.1 and 4.1 from our paper are proved in Section 3. The list of the explanatory variables considered in the real data analysis in Section 6 of our paper is provided in Section 4.

1. Mathematical justification of the approximate robust PCA

See Section 4.1 of our paper for the definition of notation. Given that the LPTN matches the normal distribution everywhere except in the tails, the limiting posterior of $(\tilde{\mathbf{Z}}_q, \tilde{\mathbf{L}}_q, \tilde{\mathbf{A}}_q, \eta)$ based on the exact robust PCA is similar to that arising from the traditional PCA model with normal errors based on $\mathbf{C}^*$, as the outliers moves away from the general trend. This means that the exact robust PCA applied to $\mathbf{C}$ leads to essentially the same singular value decomposition as a traditional PCA applied to $\mathbf{C}^*$ (in the limit). The approximate robust PCA method relies on this equivalence.

The first step in performing an approximate robust PCA is to obtain $\mathbf{C}$ by standardising the columns of the original data set. Location and scale estimates $\widehat{\mu}_j$ and $\widehat{\sigma}_j$ are thus used to standardise Column j, $j = 1, \ldots, p$. Relying on a robust location-scale model as in [2], with an LPTN error distribution and $\rho := 0.95$, ensures that $(\widehat{\mu}_j, \widehat{\sigma}_j) \longrightarrow (\widehat{\mu}_j^{-\mathcal{O}}, \widehat{\sigma}_j^{-\mathcal{O}})$, where $(\widehat{\mu}_j^{-\mathcal{O}}, \widehat{\sigma}_j^{-\mathcal{O}})$ are estimates based on nonoutliers only. Note that the robust location-scale model is the linear regression model with only the intercept. For large n, we also have $(\widehat{\mu}_j^{-\mathcal{O}}, \widehat{\sigma}_j^{-\mathcal{O}}) \approx (\widehat{\mu}_j^*, \widehat{\sigma}_j^*)$, where $(\widehat{\mu}_j^*, \widehat{\sigma}_j^*)$ are the sample mean and standard deviation obtained from $\mathbf{C}^*$, which is based on the normality of errors and in which outliers are replaced by their vertical projection. Denote by $c_{ij}^{\mathcal{O}}$ the outlying values; they are excluded for the estimation of $(\widehat{\mu}_j^{-\mathcal{O}}, \widehat{\sigma}_j^{-\mathcal{O}})$ and replaced by their vertical projection, denoted by c_{ij}^*, for the estimation of $(\widehat{\mu}_j^*, \widehat{\sigma}_j^*)$. Provided that n is large enough, the impact of those points on the sample mean and standard deviation will indeed be negligible; furthermore, it was previously argued that estimates obtained under LPTN and normal error distributions are similar. The resulting

CONTACT Philippe Gagnon. Email: philippe.gagnon@stats.ox.ac.uk

matrices $\mathbf{C}$ and $\mathbf{C}^*$ are the same in the limit, except on lines containing outliers.

The second step in performing the approximate robust PCA consists in computing robust correlations between all pairs of columns in $\mathbf{C}$. We know that the correlation between the standardised columns j_1 and j_2 of $\mathbf{C}^*$ is $\widehat{\beta}^{\mathcal{N}}_{j_1,j_2}$, the OLS slope estimate. We are interested in comparing the robust slope estimator (applied to columns of the matrix $\mathbf{C}$) to $\widehat{\beta}^{\mathcal{N}}_{j_1,j_2}$. When using a robust regression model as in [3] with an LPTN error distribution and $\rho := 0.95$, we find $\widehat{\beta}_{j_1,j_2} \longrightarrow \widehat{\beta}^{-\mathcal{O}}_{j_1,j_2}$, where $\widehat{\beta}^{-\mathcal{O}}_{j_1,j_2}$ is the robust slope estimate obtained using nonoutliers only. Again, for large n, we find $\widehat{\beta}^{-\mathcal{O}}_{j_1,j_2} \approx \widehat{\beta}^{\mathcal{N}}_{j_1,j_2}$. The robust correlation matrix obtained from $\mathbf{C}$ is thus asymptotically equal to the correlation matrix obtained from $\mathbf{C}^*$. Its diagonal elements are equal to 1; for simplicity, we set the upper diagonal entries to $\widehat{\beta}_{j_1,j_2}$, where Column j_2 plays the role of the dependent variable; we then make the matrix symmetrical.

The PCs $\widehat{\mathbf{Z}}_q$ are ultimately computed using $\mathbf{C}\widehat{\mathbf{v}}_j$, with $\widehat{\mathbf{v}}_j$ being the j-th eigenvector of the robust correlation matrix of $\mathbf{C}$.

2. Validity of our prior structure

Relying on improper priors such as $\pi(\sigma_k, \boldsymbol{\beta}_k \mid k) = c_k/\sigma_k$ may lead to inconsistencies in model selection (see [1]). For instance, one could select different constants c_k in different models so as to yield the desired conclusions. In this section, we show that the Jeffreys-Lindley paradox does not arise in our PCR context under the normal distribution assumption. It is thus expected to not arise either under the robust LPTN distribution, given its similarity to the normal.

Consider Models s and t, where Model s is nested in Model t. The ratio of the posterior probabilities of these two models is given by (see Proposition 3.1 in our paper)

$$\begin{aligned}\frac{\pi(t \mid \mathbf{y})}{\pi(s \mid \mathbf{y})} = \frac{\Gamma((n-d_s)/2-(d_t-d_s)/2)}{\Gamma((n-d_s)/2)((n-d_s)/2)^{-(d_t-d_s)/2}}\, n^{-(d_t-d_s)/2} \left(\frac{\|\mathbf{y}-\widehat{\mathbf{y}}_s\|^2/(n-1)}{\|\mathbf{y}-\widehat{\mathbf{y}}_t\|^2/(n-1)}\right)^{n/2} \\ \times \frac{\pi^{d_t/2}}{\pi^{d_s/2}} \frac{((n-d_s)/2)^{-(d_t-d_s)/2}}{n^{-(d_t-d_s)/2}} \frac{\left(\|\mathbf{y}-\widehat{\mathbf{y}}_t\|^2/(n-1)\right)^{d_t/2}}{\left(\|\mathbf{y}-\widehat{\mathbf{y}}_s\|^2/(n-1)\right)^{d_s/2}} \frac{\pi(t)}{\pi(s)}.\end{aligned} \tag{1}$$

The difference between the Bayesian information criteria (BIC, [6]) of Models t and s is given by

$$\begin{aligned}\mathrm{BIC}_t - \mathrm{BIC}_s &= n\log\left(\|\mathbf{y}-\widehat{\mathbf{y}}_t\|^2/n\right) + (d_t+1)\log n \\ &\quad - n\log\left(\|\mathbf{y}-\widehat{\mathbf{y}}_s\|^2/n\right) - (d_s+1)\log n \\ &= n\log\left(\frac{\|\mathbf{y}-\widehat{\mathbf{y}}_t\|^2/n}{\|\mathbf{y}-\widehat{\mathbf{y}}_s\|^2/n}\right) + (d_t-d_s)\log n.\end{aligned}$$

Given that the first ratio on the right hand side of (1) converges to 1 as $n \longrightarrow \infty$, we have that $\exp\{-(\mathrm{BIC}_t - \mathrm{BIC}_s)/2\}$ asymptotically behaves like the first part on the right hand side of (1). The terms $\left(\|\mathbf{y}-\widehat{\mathbf{y}}_k\|^2/(n-1)\right)^{d_k/2}$ in (1) converge towards a constant (in n) and are thus dominated. The other terms in (1) are either constant in terms of n or dominated as well. Therefore, $\pi(t|\mathbf{y})/\pi(s|\mathbf{y})$ and $\exp\{-(\mathrm{BIC}_t - \mathrm{BIC}_s)/2\}$

share the same asymptotic behaviour. This will be sufficient to prove that the prior structure does not prevent the Bayesian variable selection procedure to be consistent, in the same sense as [1]. If the "true" model is among the models considered, then its posterior probability converges to 1 as n increases. Further technical details are required for a rigorous proof. Empirical evidences also point towards the validity of our claim.

It would be interesting to investigate the asymptotic behaviour in the more general context of traditional linear regression. The fact that the regressors are standardised and linearly independent plays a role in the sketch of the proof presented above. It would however be surprising if a similar prior structure, but with slightly correlated standardised regressors, led to inconsistencies.

In practice (with finite samples), one may set the prior $\pi(k)$ to be proportional to $\pi^{-d_k/2}$ times a prior opinion about $\left(\|\mathbf{y}-\widehat{\mathbf{y}}_k\|^2/(n-1)\right)^{-d_k/2}$, to cancel the effect of these two terms in (1). In the numerical analyses, we set $\pi(k) \propto 1$ because we do not have relevant information. Note that the robust approach proposed in this paper can be used with any informative prior such as those in [5].

3. Proofs

Proof of Proposition 3.1. The proof is essentially a computation using that $f := \mathcal{N}(0,1)$ and the structure of the principal components. First,

$$\begin{aligned}\pi(k,\sigma_k,\boldsymbol{\beta}_k \mid \mathbf{y}) &\propto f(\mathbf{y} \mid k,\sigma_k,\boldsymbol{\beta}_k)\pi(\sigma_k,\boldsymbol{\beta}_k \mid k)\pi(k)\\ &\propto f(\mathbf{y} \mid k,\sigma_k,\boldsymbol{\beta}_k)(1/\sigma_k)\pi(k).\end{aligned}$$

The likelihood function for a given model is

$$\begin{aligned}f(\mathbf{y} \mid k,\sigma_k,\boldsymbol{\beta}_k) &= \prod_{i=1}^{n}\frac{1}{\sigma_k\sqrt{2\pi}}\exp\left\{-\frac{1}{2\sigma_k^2}(y_i-\mathbf{x}_{i,k}^T\boldsymbol{\beta}_k)^2\right\}\\ &= \frac{1}{\sigma_k^n(2\pi)^{n/2}}\exp\left\{-\frac{1}{2\sigma_k^2}\sum_{i=1}^{n}(y_i-\mathbf{x}_{i,k}^T\boldsymbol{\beta}_k)^2\right\}.\end{aligned}$$

We now analyse the sum in the exponential:

$$\begin{aligned}\sum_{i=1}^{n}(y_i-\mathbf{x}_{i,k}^T\boldsymbol{\beta}_k)^2 &= \sum_{i=1}^{n}y_i^2 - 2\sum_{i=1}^{n}y_i\sum_{j=1}^{d_k}x_{iI_{j,k}}\beta_{j,k} + \sum_{i=1}^{n}\left(\sum_{j=1}^{d_k}x_{iI_{j,k}}\beta_{j,k}\right)^2\\ &= n-1-2\sum_{j=1}^{d_k}\beta_{j,k}\sum_{i=1}^{n}y_ix_{iI_{j,k}} + \sum_{i=1}^{n}\left(\sum_{j=1}^{d_k}x_{iI_{j,k}}\beta_{j,k}\right)^2,\end{aligned}$$

using that $\sum_{i=1}^{n}y_i^2 = n-1$. We also have

$$\sum_{i=1}^{n}\left(\sum_{j=1}^{d_k}x_{iI_{j,k}}\beta_{j,k}\right)^2 = \sum_{i=1}^{n}\left(\sum_{j=1}^{d_k}(x_{iI_{j,k}}\beta_{j,k})^2 + \sum_{j,s=1(j\neq s)}^{d_k}x_{iI_{j,k}}\beta_{j,k}x_{iI_{s,k}}\beta_{s,k}\right)$$

$$= \sum_{j=1}^{d_k} \beta_{j,k}^2 \sum_{i=1}^{n} x_{iI_{j,k}}^2,$$

using $\sum_{i=1}^{n} x_{ij}x_{is} = 0$ for all $j, s \in \{2, \ldots, d\}$ with $j \neq s$, $x_{11} = \ldots = x_{n1} = 1$, $(1/n)\sum_{i=1}^{n} x_{ij} = 0$ for all $j \in \{2, \ldots, d\}$. Consequently,

$$\begin{aligned}\sum_{i=1}^{n}(y_i - \mathbf{x}_{i,k}^T\boldsymbol{\beta}_k)^2 &= n - 1 - 2\sum_{j=1}^{d_k}\beta_{j,k}\sum_{i=1}^{n} y_i x_{iI_{j,k}} + \sum_{j=1}^{d_k}\beta_{j,k}^2\sum_{i=1}^{n} x_{iI_{j,k}}^2 \\ &= n - 1 - \mathbb{1}(k \geq 2)2\sum_{j=2}^{d_k}\beta_{j,k}\sum_{i=1}^{n} y_i x_{iI_{j,k}} + n\beta_{1,k}^2 \\ &\quad + \mathbb{1}(k \geq 2)(n-1)\sum_{j=2}^{d^k}\beta_{j,k}^2,\end{aligned}$$

using again $x_{11} = \ldots = x_{n1} = 1$, $\sum_{i=1}^{n} y_i = 0$ and $\sum_{i=1}^{n} x_{ij}^2 = n - 1$ for all $j \in \{2, \ldots, d\}$. We also have

$$\begin{aligned}&\mathbb{1}(k \geq 2)\left((n-1)\sum_{j=2}^{d_k}\beta_{j,k}^2 - 2\sum_{j=2}^{d_k}\beta_{j,k}\sum_{i=1}^{n} y_i x_{iI_{j,k}}\right) \\ &= \mathbb{1}(k \geq 2)(n-1)\sum_{j=2}^{d_k}\left(\beta_{j,k}^2 - 2\beta_{j,k}\frac{\sum_{i=1}^{n} y_i x_{iI_{j,k}}}{n-1}\right) \\ &= \mathbb{1}(k \geq 2)(n-1)\sum_{j=2}^{d_k}\left(\beta_{j,k} - \frac{\sum_{i=1}^{n} x_{iI_{j,k}} y_i}{n-1}\right)^2 \\ &\quad - \mathbb{1}(k \geq 2)(n-1)\sum_{j=2}^{d_k}\left(\frac{\sum_{i=1}^{n} x_{iI_{j,k}} y_i}{n-1}\right)^2.\end{aligned}$$

Putting this together leads to:

$$\begin{aligned}&\pi(k, \sigma_k, \boldsymbol{\beta}_k \mid \mathbf{y}) \\ &\propto \pi(k)(2\pi)^{d_k/2}\frac{1}{\sigma_k^{n-d_k+1}}\exp\left\{-\frac{n-1}{2\sigma_k^2}\left(1 - \mathbb{1}(k \geq 2)\sum_{j \in I_k \setminus \{1\}}\left(\frac{\sum_{i=1}^{n} x_{ij} y_i}{n-1}\right)^2\right)\right\} \\ &\quad \times \frac{1}{\sigma_k\sqrt{2\pi}}\exp\left\{-\frac{n}{2\sigma_k^2}\beta_{1,k}^2\right\} \\ &\quad \times \left(\mathbb{1}(k=1) + \mathbb{1}(k \geq 2)\prod_{j=2}^{d_k}\frac{1}{\sigma_k\sqrt{2\pi}}\exp\left\{-\frac{n-1}{2\sigma_k^2}\left(\beta_{j,k} - \frac{\sum_{i=1}^{n} x_{iI_{j,k}} y_i}{n-1}\right)^2\right\}\right).\end{aligned}$$

We multiply and divide by the appropriate terms. The only remaining thing to show is that

$$n-1\left(1-\mathbb{1}(k\geq 2)\sum_{j\in I_k\setminus\{1\}}\left(\frac{\sum_{i=1}^n x_{ij}y_i}{n-1}\right)^2\right)=\|\mathbf{y}-\widehat{\mathbf{y}}_k\|_2^2.$$

Firstly, $n-1=\|\mathbf{y}\|_2^2$. Also,

$$\begin{aligned}\|\mathbf{y}\|_2^2&=\|\mathbf{y}-\widehat{\mathbf{y}}_k+\widehat{\mathbf{y}}_k\|_2^2\\&=\|\mathbf{y}-\widehat{\mathbf{y}}_k\|_2^2+(\mathbf{y}-\widehat{\mathbf{y}}_k)^T\widehat{\mathbf{y}}_k+\widehat{\mathbf{y}}_k^T(\mathbf{y}-\widehat{\mathbf{y}}_k)+\widehat{\mathbf{y}}_k^T\widehat{\mathbf{y}}_k.\end{aligned}$$

We know that $(\mathbf{y}-\widehat{\mathbf{y}}_k)^T\widehat{\mathbf{y}}_k=\widehat{\mathbf{y}}_k^T(\mathbf{y}-\widehat{\mathbf{y}}_k)=0$ because $\mathbf{y}-\widehat{\mathbf{y}}_k$ is the vector of residuals which is orthogonal to $\widehat{\mathbf{y}}_k$. Finally,

$$\begin{aligned}\widehat{\mathbf{y}}_k^T\widehat{\mathbf{y}}_k=(\mathbf{X}_k\widehat{\boldsymbol{\beta}}_k)^T\mathbf{X}_k\widehat{\boldsymbol{\beta}}_k=\widehat{\boldsymbol{\beta}}_k^T\mathbf{X}_k^T\mathbf{X}_k\widehat{\boldsymbol{\beta}}_k&=(n-1)\|\widehat{\boldsymbol{\beta}}_k\|_2^2\\&=(n-1)\mathbb{1}(k\geq 2)\sum_{j\in I_k\setminus\{1\}}\left(\frac{\sum_{i=1}^n x_{ij}y_i}{n-1}\right)^2,\end{aligned}$$

where $\mathbf{X}_k$ is the design matrix associated with Model k. □

Proof of Proposition 2.2. As explained in [4], it suffices to separately verify that the probability to go from a set A to a set B is equal to the probability to go from B to A when updating the parameters and when switching models, for accepted movements and for any appropriate A, B.

When updating the parameters, the probability to go from a set A to a set B is given by

$$\begin{aligned}\int_A \pi(k,\sigma_k,\boldsymbol{\beta}_k\mid\mathbf{y})g(1)\int_B\prod_{i=1}^{1+d_k}\varphi_i(w_i\mid k,(\sigma_k,\boldsymbol{\beta}_k)_i,\ell_k)\\\times\left(1\wedge\frac{(1/w_1)f(\mathbf{y}\mid k,\mathbf{w}_k)}{(1/\sigma_k)f(\mathbf{y}\mid k,\sigma_k,\boldsymbol{\beta}_k)}\right)d\mathbf{w}_k\,d(\sigma_k,\boldsymbol{\beta}_k).\end{aligned}$$

Using Fubini's theorem, this probability is equal to

$$\begin{aligned}\int_B \pi(k,\mathbf{w}_k\mid\mathbf{y})g(1)\int_A\prod_{i=1}^{1+d_k}\varphi_i((\sigma_k,\boldsymbol{\beta}_k)_i\mid k,w_i,\ell_k)\\\times\left(1\wedge\frac{(1/\sigma_k)f(\mathbf{y}\mid k,\sigma_k,\boldsymbol{\beta}_k)}{(1/w_1)f(\mathbf{y}\mid k,\mathbf{w}_k)}\right)d(\sigma_k,\boldsymbol{\beta}_k)\,d\mathbf{w}_k,\end{aligned}$$

which is the probability to go from B to A. Note that this is valid for all $k\in\{1,\ldots,\mathrm{K}_{\max}\}$.

The probability to switch from Model $k\in\{1,\ldots,\mathrm{K}_{\max}-1\}$, where the parameters are in the set A, to Model $k+1$, where the parameters are in the set $A'\times B$ (the set

A' is a modified version of A to account for the addition of $\mathbf{c}_{k+1}$), is given by

$$\int_A \pi(k, \sigma_k, \boldsymbol{\beta}_k \mid \mathbf{y}) g(2) \int_B q_{k+1}(u_{k+1}) \\ \times \left(1 \wedge \frac{\pi(k+1) f(\mathbf{y} \mid k+1, (\sigma_k, \boldsymbol{\beta}_k) + \mathbf{c}_{k+1}, u_{k+1})}{\pi(k) f(\mathbf{y} \mid k, \sigma_k, \boldsymbol{\beta}_k) q_{k+1}(u_{k+1})}\right) du_{k+1}\, d(\sigma_k, \boldsymbol{\beta}_k).$$

After the change of variables $(\sigma_{k+1}, \boldsymbol{\beta}_{k+1}) = ((\sigma_k, \boldsymbol{\beta}_k) + \mathbf{c}_{k+1}, u_{k+1})$, we have

$$\int_{A' \times B} \pi(k, (\sigma_{k+1}, \boldsymbol{\beta}_{k+1}^-) - \mathbf{c}_{k+1} \mid \mathbf{y}) g(2) q_{k+1}(\beta_{d_{k+1}, k+1}) \\ \times \left(1 \wedge \frac{\pi(k+1) f(\mathbf{y} \mid k+1, \sigma_{k+1}, \boldsymbol{\beta}_{k+1})}{\pi(k) f(\mathbf{y} \mid k, (\sigma_{k+1}, \boldsymbol{\beta}_{k+1}^-) - \mathbf{c}_{k+1}) q_{k+1}(\beta_{d_{k+1}, k+1})}\right) d(\sigma_{k+1}, \boldsymbol{\beta}_{k+1}).$$

This last probability is equal to

$$\int_{A' \times B} \pi(k+1, \sigma_{k+1}, \boldsymbol{\beta}_{k+1} \mid \mathbf{y}) g(3) \\ \times \left(1 \wedge \frac{\pi(k) f(\mathbf{y} \mid k, (\sigma_{k+1}, \boldsymbol{\beta}_{k+1}^-) - \mathbf{c}_{k+1}) q_{k+1}(\beta_{d_{k+1}, k+1})}{\pi(k+1) f(\mathbf{y} \mid k+1, \sigma_{k+1}, \boldsymbol{\beta}_{k+1})}\right) d(\sigma_{k+1}, \boldsymbol{\beta}_{k+1}),$$

which is the probability to switch from Model $k+1$, where the parameters are in the set $A' \times B$, to Model k, where the parameters are in the set A.

Therefore, the Markov chain $\{(K, \sigma_K, \boldsymbol{\beta}_K)(m) : m \in \mathbb{N}\}$ satisfies the reversibility condition with respect to the posterior. □

4. List of the explanatory variables used in Section 6

Name	Ticker symbol
Artis Real Estate Investment Trust	AX-UN.TO
Asanko Gold Inc.	AKG.TO
Bonterra Energy Corp.	BNE.TO
Canadian Imperial Bank Of Commerce	CM.TO
CI Financial Corp.	CIX.TO
Celestica Inc. Subordinate Voting Shares	CLS.TO
DHX Media Ltd.	DHX-B.TO
Dominion Diamond Corporation	DDC.TO
Gildan Activewear Inc.	GIL.TO
Husky Energy Inc.	HSE.TO
iPath Bloomberg Sugar Subindex	SGG
iShares MSCI Japan	EWJ
iShares 20+ Year Treasury Bond	TLT
Laurentian Bank of Canada	LB.TO
Parkland Fuel Corporation	PKI.TO
United States Oil Fund LP	USO
Vermilion Energy Inc.	VET.TO
Volume of the S&P 500	N/A

Table 1. Names of the companies, funds, and financial indicators used as explanatory variables in the analysis in Section 6 of our paper, with their ticker symbol (if available)

References

[1] G. Casella, F.J. Giròn, M.L. Martínez, and E. Moreno, *Consistency of Bayesian procedures for variable selection*, Ann. Statist. 37 (2009), pp. 1207–1228.

[2] A. Desgagné, *Robustness to outliers in location–scale parameter model using log-regularly varying distributions*, Ann. Statist. 43 (2015), pp. 1568–1595.

[3] P. Gagnon, A. Desgagné, and M. Bédard, *A new bayesian approach to robustness against outliers in linear regression*, Bayesian Anal. (2018). Advance publication.

[4] P.J. Green, *Reversible jump Markov chain Monte Carlo computation and Bayesian model determination*, Biometrika 82 (1995), pp. 711–732.

[5] A.E. Raftery, D. Madigan, and J.A. Hoeting, *Bayesian model averaging for linear regression models*, J. Amer. Statist. Assoc. 92 (1997), pp. 179–191.

[6] G. Schwarz, *Estimating the dimension of a model*, Ann. Statist. 6 (1978), pp. 461–464.