Inference on covariance matrices and operators using concentration inequalities

Adam B Kashlak[*1]

[1]*Cambridge Centre for Analysis, University of Cambridge, UK*

In the modern era of high and infinite dimensional data, classical statistical methodology is often rendered inefficient and ineffective when confronted with such big data problems as arise in genomics, medical imaging, speech analysis, and many other areas of research. Many problems manifest when the practitioner is required to take into account the covariance structure of the data during his or her analysis, which takes on the form of either a high dimensional low rank matrix or a finite dimensional representation of an infinite dimensional operator acting on some underlying function space. Thus, we propose using tools from the concentration of measure literature to construct rigorous descriptive and inferential statistical methodology for covariance matrices and operators. A variety of concentration inequalities are considered, which allow for the construction of nonasymptotic dimension-free confidence sets for the unknown matrices and operators. Given such confidence sets a wide range of estimation and inferential procedures can be and are subsequently developed.

Keywords: Sparse Matrix, Functional Data Analysis, Log Sobolev Inequality, Talagrand's Inequality, Confidence Sets

1 Overview

Concentration inequalities are a general category of results from geometry, functional analysis, and probability theory that control the tail behaviour of probability measures. In recent years, they have proved invaluable to statisticians due to their non-asymptotic dimension-free properties, which makes them particularly suitable for estimation and inference on finite samples of data living high or infinite dimensional space. Overviews of such results can be found in the monographs [3, 8, 11]. This manuscript introduces some of the author's doctoral research into using concentration inequalities for statistical estimation and inference on covariance matrices and operators.

*Corresponding author: ak852@cam.ac.uk or kashlak@ualberta.ca

1.1 Definitions and notation

Definition 1 (Empirical Covariance Matrix)**.** Let $X_1, \ldots, X_n \in \mathbb{R}^d$ be iid realizations of some random variable $X \in \mathbb{R}^d$ with unknown covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$. Then, the sample or empirical estimate for Σ is

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})^{\mathrm{T}}$$

where $\bar{X} = n^{-1} \sum_{i=1}^{n} X_i$ is the sample mean of the data.

Definition 2 (Empirical Covariance Operator)**.** For $I \subseteq \mathbb{R}$, let $f_1, \ldots, f_n \in L^2(I)$ be iid realizations of some random function $f \in L^2(I)$ with unknown covariance operator $\Sigma \in Op(L^2)$. Then, the sample or empirical estimate for Σ is

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (f_i - \bar{f}) \otimes (f_i - \bar{f}) = \frac{1}{n} \sum_{i=1}^{n} (f_i - \bar{f})^{\otimes 2} = \frac{1}{n} \sum_{i=1}^{n} \big\langle (f_i - \bar{f}), \cdot \big\rangle \, (f_i - \bar{f})$$

where $\bar{f} = n^{-1} \sum_{i=1}^{n} f_i$ is the sample mean of the data.

Definition 3 (p-Schatten norm for matrices)**.** For an arbitrary matrix $\Sigma \in \mathbb{R}^{k \times l}$ and $p \in (1, \infty)$, the p-Schatten norm is

$$\|\Sigma\|_p^p = \mathrm{tr}\left((\Sigma^{\mathrm{T}} \Sigma)^{p/2} \right) = \|\boldsymbol{\nu}\|_{\ell^p}^p = \sum_{i=1}^{\min\{k,l\}} \nu_i^p$$

where $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_{\min\{k,l\}})$ is the vector of singular values of Σ and where $\|\cdot\|_{\ell^p}$ is the standard ℓ^p norm in $\mathbb{R}^d$. In the covariance matrix case where $\Sigma \in \mathbb{R}^{d \times d}$ is symmetric and positive definite, $\|\Sigma\|_p^p = \mathrm{tr}\,(\Sigma^p) = \|\boldsymbol{\lambda}\|_{\ell^p}^p$ where $\boldsymbol{\lambda}$ is the vector of eigenvalues of Σ.

When $p = \infty$, we have the standard operator norm on Euclidean space

$$\|\Sigma\|_\infty = \sup_{v \in \mathbb{R}^d, \|v\|_{\ell^2} = 1} \|\Sigma v\|_{\ell^2} = \sup_{v \in \mathbb{R}^d, \|v\|_{\ell^2} = 1} v^{\mathrm{T}} \Sigma v.$$

For covariance matrices, this coincides with the maximal eigenvalue of Σ.

Definition 4 (p-Schatten norm for operators)**.** Given two separable Hilbert spaces H_1 and H_2, a bounded linear operator $\Sigma : H_1 \to H_2$, and some $p \in [1, \infty)$, then the p-Schatten norm is $\|\Sigma\|_p^p = \mathrm{tr}\left((\Sigma^* \Sigma)^{p/2} \right)$. For $p = \infty$, the Schatten norm is the operator norm: $\|\Sigma\|_\infty = \sup_{f \in H_1} (\|\Sigma f\|_{H_2} / \|f\|_{H_1})$. In the case that Σ is compact, self-adjoint, and trace-class, then given the associated

eigenvalues $\{\lambda_i\}_{i=1}^{\infty}$, the p-Schatten norm coincides with the standard ℓ^p norm of the eigenvalues:

$$\|\Sigma\|_p^p = \begin{cases} \|\lambda\|_{\ell^p}^p = \sum_{i=1}^{\infty} |\lambda_i|^p, & p \in [1, \infty) \\ \max_{i \in \mathbb{N}} |\lambda_i|, & p = \infty \end{cases}$$

2 Covariance matrices

Given $X_1, \ldots, X_n \in \mathbb{R}^d$, past studies have shown that the empirical estimate for the covariance matrix, Definition 1, is a very poor estimator when the underlying true Σ is high dimensional, $d \gg n$, and sparse meaning that most of the off-diagonal entries are zero or negligible. Hence, much research has gone into better estimation techniques [1, 2, 5, 15, 14]. In [10], we propose using concentration inequalities to construct a non-asymptotic confidence set for the empirical estimate and then search the confidence set in order to find an improved estimator.

Let $d(\cdot, \cdot)$ be some metric measuring the distance between two covariance matrices, and let $\psi : \mathbb{R} \to \mathbb{R}$ be monotonically increasing. Then, the general form of the concentration inequalities is

$$\mathrm{P}\left(d(\Sigma_0, \hat{\Sigma}^{\mathrm{emp}}) \geq \mathrm{E} d(\Sigma_0, \hat{\Sigma}^{\mathrm{emp}}) + r \right) \leq \mathrm{e}^{-\psi(r)},$$

which is a bound on the tail of the distribution of $d(\Sigma_0, \hat{\Sigma}^{\mathrm{emp}})$ as it deviates above its mean. Thus, to construct a $(1 - \alpha)$-confidence set, the variable $r = r_\alpha$ is chosen such that $\exp(-\psi(r_\alpha)) = \alpha$. Then, choose a $\hat{\Sigma}^{\mathrm{sp}}$ such that $d(\hat{\Sigma}^{\mathrm{sp}}, \hat{\Sigma}^{\mathrm{emp}}) \leq r_\alpha$.

The proposed search procedure is to sequentially set to zero the smallest entries in $\hat{\Sigma}^{\mathrm{emp}}$ while remaining inside the r_α-ball. The metric used is $d(\Sigma_0, \hat{\Sigma}^{\mathrm{emp}}) = \|\Sigma_0 - \hat{\Sigma}^{\mathrm{emp}}\|_p^{1/2}$ where $\|\cdot\|_p$ is the p-Schatten norm from Definition 3. This metric is shown to be Lipschitz $n^{-1/2}$ with respect to Euclidean distance in $\mathbb{R}^{d \times n}$.

In [10], three types of distributional assumptions are considered: log concave measures; sub-exponential measures; bounded random variables. In summary, applying our methodology to log concave measures, which include the multivariate Gaussian distribution, yielded excellent theoretical and experimental results. Our method is particularly good at support recovery or "sparsistency" in this case. For sub-exponential measures, the concentration inequalities do not yield nice theoretical results, but the methodology still gives good performance in simulation studies. This approach fails in the bounded random variable case as the resulting confidence sets are not dimension-free.

3 Covariance operators

In the functional data setting, $f_1, \ldots, f_n \in L^2(I)$ are iid random functions with $I \subseteq \mathbb{R}$. Similarly to the high dimensional case, covariance operators are of critical importance to inference and hypothesis testing. For example, the development of k-sample tests for the equality of covariance is a major area of research [4, 7, 12, 13].

In [9], we propose our own k-sample test for the equality of covariance by first using Talagrand's concentration inequality [16] in the Banach space setting to construct confidence sets for each of the covariance operator. For some desired p-Schatten norm, Definition 4, $\|\cdot\|_p$, with $p \in [1, \infty)$ and with conjugate $q = p/(p-1)$, we require the following terms, which correspond to the distance between the empirical covariance estimate and the true covariance operator and a weak variance term for this random variable:

$$Z = \left\| \frac{1}{n} \sum_{i=1}^n f_i \otimes f_i - \mathrm{E}\left(f_i \otimes f_i\right) \right\|_p = \left\| \hat{\Sigma} - \Sigma \right\|_p$$

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n \sup_{\|\Pi\|_q \leq 1} \mathrm{E} \left\langle f_i^{\otimes 2} - \mathrm{E} f_i^{\otimes 2}, \Pi \right\rangle^2.$$

In the above equation, the supremum is to be taken over a countably dense subset of the unit ball of $\Pi \in Op(L^2)$. For some $U \geq \|f_i^{\otimes 2}\|_{L^2}^2$ and $v_n = 2U\mathrm{E}Z + n\sigma^2$, the initial level $(1-\alpha)$ confidence set constructed is

$$C_{n,1-\alpha} = \left\{ \Sigma : Z \leq \mathrm{E}Z + \sqrt{-2v_n \log(2\alpha)/n} - U\log(2\alpha)/(3n) \right\}.$$

To make this confidence set usable on real data, the Rademacher average $R_n = n^{-1} \sum_{i=1}^n \varepsilon_i ((f_i - \bar{f})^{\otimes 2} - \hat{\Sigma})$, where $\mathrm{P}(\varepsilon_i = 1) = \mathrm{P}(\varepsilon_i = -1) = 0.5$ will be used as a proxy for the unknown $\mathrm{E}Z$.

In [9], this is not only applied to k-sample tests for equality of covariance, but also to the classification and clustering of functional data. This methodology is applied to a set of phoneme data detailed in [6], which is a collection of 400 log-periodograms for each of five different phonemes: /ɑ/ as in the vowel of "dark"; /ɔ/ as in the first vowel of "water"; /d/ as in the plosive of "dark"; /i/ as in the vowel of "she"; /ʃ/ as in the fricative of "she". Each curve contains the first 150 frequencies from a 32 ms sound clip sampled at a rate of 16-kHz. Comparisons of our concentration-based methodology with other methods of classification and clustering can be found in Tables 1 and 2, respectively.

	/ɑ/	/ɔ/	/d/	/i/	/ʃ/
CoM	76.9	76.8	96.6	**98.5**	99.4
KNN	72.4	79.1	**98.5**	97.4	**100.**
Kernel	72.0	**80.5**	98.4	97.2	99.9
GLM	**79.0**	72.3	98.2	95.9	99.2
Tree	70.8	69.4	95.6	87.8	92.6

Table 1: Percentage of correct classification of the five phonemes against the five methods: our concentration of measure approach (CoM); k-nearest-neighbours (KNN); kernel method (Kernel); generalized linear model (GLM); and regression trees (Tree).

	Concentration					**k-means**				
Cluster	A	B	C	D	E	A	B	C	D	E
/ɑ/	281	119	0	0	0	281	119	0	0	0
/ɔ/	125	273	1	1	0	126	272	1	1	0
/d/	0	0	384	15	1	0	2	386	10	2
/i/	1	0	1	393	5	1	3	2	381	13
/ʃ/	0	0	0	3	397	0	0	0	2	398

Table 2: Clustering 2000 phoneme curves into 5 clusters. Similar results achieved by both the concentration and k-means methods.

Acknowledgements: The author would like to acknowledge the support of his thesis advisers Professors John AD Aston and Richard Nickl from the University of Cambridge Statistical Laboratory. He would also like to thank Professor Linglong Kong at the University of Alberta for his collaboration on the research contained in Section 2.

References

[1] P. J. Bickel and E. Levina. Covariance regularization by thresholding. *The Annals of Statistics*, pages 2577–2604, 2008.

[2] J. Bien and R. J. Tibshirani. Sparse estimation of a covariance matrix. *Biometrika*, 98(4):807–820, 2011.

[3] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.

[4] A. Cabassi, D. Pigoli, P. Secchi, and P. A. Carter. Permutation tests for the equality of covariance operators of functional data with applications to evolutionary biology. *arXiv preprint arXiv:1701.05870*, 2017.

[5] T. Cai and W. Liu. Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association*, 106(494):672–684, 2011.

[6] F. Ferraty and P. Vieu. Curves discrimination: a nonparametric functional approach. *Computational Statistics & Data Analysis*, 44(1):161–173, 2003.

[7] S. Fremdt, J. G. Steinebach, L. Horváth, and P. Kokoszka. Testing the equality of covariance operators in functional samples. *Scandinavian Journal of Statistics*, 40(1):138–152, 2013.

[8] E. Giné and R. Nickl. *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge University Press, 2016.

[9] A. B. Kashlak, J. A. D. Aston, and R. Nickl. Inference on covariance operators via concentration inequalities: k-sample tests, classification, and clustering via Rademacher complexities. *arXiv preprint arXiv:1604.06310*, 2016.

[10] A. B. Kashlak and L. Kong. A concentration inequality based methodology for sparse covariance estimation. *arXiv preprint arXiv:1705.02679*, 2017.

[11] M. Ledoux. *The concentration of measure phenomenon*, volume 89. American Mathematical Soc., 2001.

[12] V. M. Panaretos, D. Kraus, and J. H. Maddocks. Second-order comparison of Gaussian random functions and the geometry of DNA minicircles. *Journal of the American Statistical Association*, 105(490):670–682, 2010.

[13] D. Pigoli, J. A. Aston, I. L. Dryden, and P. Secchi. Distances and inference for covariance operators. *Biometrika*, page asu008, 2014.

[14] A. J. Rothman. Positive definite estimators of large covariance matrices. *Biometrika*, 99(3):733–740, 2012.

[15] A. J. Rothman, E. Levina, and J. Zhu. Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association*, 104(485):177–186, 2009.

[16] M. Talagrand. New concentration inequalities in product spaces. *Inventiones mathematicae*, 126(3):505–563, 1996.