SPCAvRP 论文阅读与思考

论文题目:
Sparse principal component analysis via random projections

SPCAvRP 论文阅读与思考

1. 论文泛读
    1.1. 标题
    1.2. 摘要
        1.2.1. 方案
        1.2.2. 优势
        1.2.3. 要点
    1.3. 介绍
        1.3.1. 前人方法
            1.3.1.1. PCA
            1.3.1.2. 高维PCA
            1.3.1.3. 迭代算法的弊端
        1.3.2. SPCAvRP
    1.4. 小标题
    1.5. 结论
    1.6. 图表
        1.6.1. 模型的参数选择【4.2】
        1.6.2. 与他人模型比较
        1.6.3. Microarray数据
2. 论文精读
    2.1. SPCA via random projections【第二章】
        2.1.1. 单一主成分估计
        2.1.2. 复合主成分估计
    2.2. Numerical experiments【第四章】
3. 实验要点与思考

1. 论文泛读

1.1. 标题

通过随机矩阵投影来进行稀疏PCA

1.2. 摘要

1.2.1. 方案

- 通过样本协方差矩阵选出多个随机投影矩阵
- 使用这些随机投影矩阵获取多个特征向量信息
- 在这些特征向量信息上进行SPCA

1.2.2. 优势

- 和其他方法不同，该方案不需要迭代，因此初始值选差了结果也差不到哪去
- 数值实验确认了该算法在有限样本上具有高度竞争性【表现很强】

1.2.3. 要点

- 平衡计算量与统计
- 为了达到目的，有效样本大小与随机矩阵投影的互相影响
- minmax:minimizing the possible loss for a worst case (maximum loss) scenario.
- When dealing with gains, it is referred to as “maximin”—to maximize the minimum gain

1.3. 介绍

1.3.1. 前人方法

1.3.1.1. PCA

- 方法：

1. 找出母体中贡献最多方差的方向，将数据沿着该方向投影。
2. 想估算出该方向，就计算样本协方差矩阵的主要特征向量

- 缺点：数据维度p和数据容量n一样大，就容易崩

as the sample size n. More precisely, suppose that $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N_p(0, \Sigma)$ are observations from a Gaussian distribution with a spiked covariance matrix $\Sigma = I_p + v_1 v_1^\top$ whose leading eigenvector is $v_1 \in \mathcal{S}^{p-1} := \{v \in \mathbb{R}^p : \|v\| = 1\}$, and let $\hat{v}_1$ denote the leading unit-length eigenvector of the sample covariance matrix $\hat{\Sigma} := n^{-1} \sum_{i=1}^n X_i X_i^\top$. Then Johnstone and Lu

(2009) and Paul (2007) showed that $\hat{v}_1$ is a consistent estimator of v_1, i.e. $|\hat{v}_1^\top v_1| \xrightarrow{p} 1$, if and only if $p = p_n$ satisfies $p/n \to 0$ as $n \to \infty$. It is also worth noting that the principal component v_1 may be a linear combination of all elements of the canonical basis in $\mathbb{R}^p$, which can often make it difficult to interpret the estimated projected directions (Jolliffe, Trendafilov and Uddin, 2003).

1. 啊在这里强调下，向量/矩阵范数默认是二范数，开过根号的那种。
2. v_1的估计是除以了n的；因为这玩意儿估计的v就是大小为1

1.3.1.2. 高维PCA

To remedy this situation, and to provide additional interpretability to the principal components in high-dimensional settings, Jolliffe, Trendafilov and Uddin (2003) and Zou, Hastie and Tibshirani (2006) proposed Sparse Principal Component Analysis (SPCA). Here it is assumed that the leading population eigenvectors belong to the k-sparse unit ball

$$\mathcal{B}_0^{p-1}(k) := \left\{ v = (v^{(1)}, \dots, v^{(p)})^\top \in \mathcal{S}^{p-1} : \sum_{j=1}^{p} \mathbb{1}_{\{v^{(j)} \neq 0\}} \leq k \right\}$$

for some $k \in \{1, \dots, p\}$. In addition to the easier interpretability, a great deal of research effort has shown that such an assumption facilitates improved estimation performance (e.g. Johnstone and Lu, 2009; Paul and Johnstone, 2012; Vu and Lei, 2013; Cai, Ma and Wu, 2013; Ma, 2013; Wang, Berthet and Samworth, 2016a). To give a flavor of these results, let $\mathcal{V}_n$ denote the set of all estimators of v_1, i.e. the class of Borel measurable functions from $\mathbb{R}^{n\times p}$ to $\mathcal{S}^{p-1}$. Vu and Lei (2013) introduce a class $\mathcal{Q}$ of sub-Gaussian distributions whose first principal component v_1 belongs to $\mathcal{B}_0^{p-1}(k)$ and show that

$$\inf_{\tilde{v}_1 \in \mathcal{V}_n} \sup_{Q \in \mathcal{Q}} \mathbb{E}_Q\{1 - (\tilde{v}_1^\top v_1)^2\} \asymp \frac{k \log p}{n}. \tag{1}$$

Here, $a_n \asymp b_n$ means $0 < \liminf_{n\to\infty} |a_n/b_n| \leq \limsup_{n\to\infty} |a_n/b_n| < \infty$. Thus, consistent estimation is possible in this framework provided only that $k = k_n$ and $p = p_n$ satisfy $(k \log p)/n \to 0$. Vu and Lei (2013) show further that this estimation rate is achieved by the natural estimator

$$\hat{v}_1 \in \underset{v \in \mathcal{B}_0^{p-1}(k)}{\operatorname{argmax}}\, v^\top \hat{\Sigma} v. \tag{2}$$

However, results such as (1) do not complete the story of SPCA. Indeed, computing the estimator defined in (2) turns out to be an NP-hard problem (e.g. Tillmann and Pfetsch,

- 请注意，（2）式的带上标的Σ不是求和符号
- 为了解决（2）式中的非凸优化问题，前人提出了L-1惩罚项的方案。然而虽然能提高速度，但是没理论支持（作者后面会笑的）
- 为了解决（2）也有使用半正定松弛法的，但是慢（作者后面会笑的）
- 重点！ it is now understood that, conditional on a Planted Clique hypothesis from theoretical computer science, there is an **asymptotic regime** in which **no randomized polynomial time algorithm can attain the minimax optimal rate**
（ref:Wang, T., Berthet, Q. and Samworth, R. J. (2016a) Statistical and computational trade-offs in estimation of sparse principal components. Ann. Statist., 44, 1896–1930.
)

1.3.1.3. 迭代算法的弊端

- 迭代的算法在确定的情况下，初值与真实值对应得很好：
Various fast, iterative algorithms were introduced by Johnstone and Lu (2009), Paul and Johnstone (2012), and Ma (2013); these have been shown to attain the minimax rate under certain conditions, provided that the initial starting point is reasonably well-aligned with the true signal.
- the loss function：

$$L(u, v) := \sin \measuredangle(u, v) = \{1 - (u^\top v)^2\}^{1/2}, \tag{3}$$

图表的是做100次取平均

- 迭代算法：初值不好全完蛋。 Remarkably, each of the previously proposed algorithms we tested produces estimates that are almost orthogonal to the true principal component！啧，这感叹号用的，我怀疑作者在笑。笑啥，你用的是人家程序的默认初始化程序……等等，好像确实可以笑人家2333

1.3.2. SPCAvRP

- 作者还是很满意的，our algorithm, which we refer to as SPCAvRP
and implement in a publicly available R package ，is also attractive for both theoretical and computational reasons。对比前人要么没理论支持要么慢的情况，确实是很不错的trade-off
- 当有效样本量很大的时候，想要得到目标结果，只需要随着样本维度p的增大稍快增加PR数目
- 但这可没违背2016a那篇文章， which applies to an intermediate
effective sample size regime where the SPCAvRP algorithm would require an exponential
number of projections to attain the optimal rate.令人尴尬的是（？）作者的算法是并行的（林源：拍桌笑），并且不用计算Σ的估计值，因为用RP提取目标子矩阵来算就够了（2016a:?）。当维度p非常大的时候，能显著节省计算。在section 4中提到，有使用数值实验和真实数据的有限样本估计来进行算法比较
- 本文也参考了贪婪算法：We also mention the computationallyefficient combinatorial approaches proposed by Moghaddam, Weiss and Avidan (2006) and
d'Aspremont, Bach and El Ghaoui (2008) that aim to find solutions to the optimization
problem in (2) using greedy methods.
- PR应用挺广的，也有例子：

Algorithms based on random projections have recently been shown to be highly effective for several different problems in high-dimensional statistical inference. For instance, in the context of high-dimensional classification, Cannings and Samworth (2017) showed that their random projection ensemble classifier that aggregates over projections that yield small estimates of the test error can result in excellent performance. Marzetta, Tucci and Simon (2011) employ an ensemble of random projections to construct an estimator of the population covariance matrix and its inverse in the setting where $n < p$. Fowler (2009) introduced a so-called compressive-projection PCA that reconstructs the sample principal components

from many low-dimensional projections of the data. Finally, to decrease the computational burden of classical PCA, Qi and Hughes (2012) and Pourkamali-Anaraki and Hughes (2014) propose estimating $v_1(\Sigma)$ by the leading eigenvector of $n^{-1}\sum_{i=1}^{n} P_i X_i X_i^\top P_i$, where $P_1, \ldots, P_n$ are random projections of a particular form.

- 记号：

Notation. We conclude this introduction with some notation used throughout the paper. For a vector $u \in \mathbb{R}^p$, we write $u^{(j)}$ for its jth component and let $\|u\| := \{\sum_{j=1}^{p}(u^{(j)})^2\}^{1/2}$ denote its Euclidean norm. For a real symmetric matrix $U \in \mathbb{R}^{p\times p}$, we let $\lambda_1(U) \geq \lambda_2(U) \geq \ldots \geq \lambda_p(U)$ denote its eigenvalues, arranged in decreasing order. In addition, we define the leading eigenvector of U by

$$v_1(U) := \operatorname*{sargmax}_{v\in\mathcal{S}^{p-1}} v^\top U v,$$

where sargmax denotes the smallest element of the argmax in the lexicographic ordering. In the special case where $U = \Sigma$, we drop the argument, and write the eigenvalues and eigenvectors as $\lambda_r = \lambda_r(\Sigma)$ and $v_r = v_r(\Sigma)$, respectively. We also define $U^{(j,j')}$ to be the (j,j')th entry of U, and write $\|U\|_{\mathrm{op}} := \sup_{\|x\|=1}\|Ux\|$ for the operator norm of matrix U.

For $r \in \mathbb{N}$, let $[r] := \{1,\ldots,r\}$. Let

$$S_1 = S_1(v_1) := \{j \in [p] : v_1^{(j)} \neq 0\}$$

denote the support of the vector $v_1 \in \mathbb{R}^p$. We write $v_1^{\min} := \min_{j\in S_1}|v_1^{(j)}|$ for the smallest non-zero component of v_1 in absolute value.

For any index subset $S \subseteq [p]$ we write P_S to denote the projection onto the span of $\{e_j : j \in S\}$, where $e_1,\ldots,e_p$ are the standard Euclidean basis vectors in $\mathbb{R}^p$, so that P_S is a $p \times p$ diagonal matrix whose jth diagonal entry is $\mathbb{1}_{\{j\in S\}}$. Finally, for $a, b \in \mathbb{R}$, we write $a \lesssim b$ to mean that there exists a universal constant $C > 0$ such that $a \leq Cb$.

请注意这里：用S_1记录了v_1的非零元素的index。也记录了v_1的绝对值最小的非零元素的index。矩阵P_S是对角线上元素为1的p*p大小的矩阵

1.4. 小标题

1 Introduction
- 2 SPCA via random projections
  - 2.1 Single principal component estimation
  - 2.2 Multiple principal component estimation
- 3 Theoretical guarantees
  - 3.1 Single-spiked model with homogeneous signal
  - 3.2 General distributions
- 4 Numerical experiments
  - 4.1 Dependence of risk on problem parameters
  - 4.2 Choice of input parameters
    - 4.2.1 Choice of A and B
    - 4.2.2 Choice of d and l
  - 4.3 Comparison with existing methods
    - 4.3.1 First principal component
    - 4.3.2 Higher-order components
    - 4.3.3 Microarray data

A Proofs of theoretical results

第三章和附录先直接跳过[捂脸]

1.5. 结论

我手上这版是没有结论和讨论的，arXiv上的进度看了也没有。只能说果然大佬。

1.6. 图表

第一章的是举例。第二章的图（？）主要是算法。第三章先跳过。第四章是数值实验

1.6.1. 模型的参数选择【4.2】

1. p,A,B:

- 这里d,l,k都是固定的。变的是p,A,B

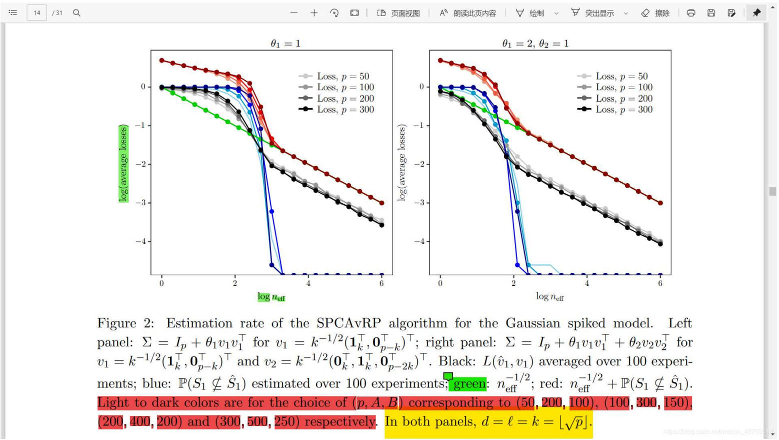

Figure 2: Estimation rate of the SPCAvRP algorithm for the Gaussian spiked model. Left panel: $\Sigma = I_p + \theta_1 v_1 v_1^\top$ for $v_1 = k^{-1/2}(\mathbf{1}_k^\top, \mathbf{0}_{p-k}^\top)^\top$; right panel: $\Sigma = I_p + \theta_1 v_1 v_1^\top + \theta_2 v_2 v_2^\top$ for $v_1 = k^{-1/2}(\mathbf{1}_k^\top, \mathbf{0}_{p-k}^\top)^\top$ and $v_2 = k^{-1/2}(\mathbf{0}_k^\top, \mathbf{1}_k^\top, \mathbf{0}_{p-2k}^\top)^\top$. Black: $L(\hat{v}_1, v_1)$ averaged over 100 experiments; blue: $\mathbb{P}(S_1 \not\subseteq \hat{S}_1)$ estimated over 100 experiments; green: $n_{\mathrm{eff}}^{-1/2}$; red: $n_{\mathrm{eff}}^{-1/2} + \mathbb{P}(S_1 \not\subseteq \hat{S}_1)$. Light to dark colors are for the choice of (p, A, B) corresponding to $(50, 200, 100)$, $(100, 300, 150)$, $(200, 400, 200)$ and $(300, 500, 250)$ respectively. In both panels, $d = \ell = k = \lfloor\sqrt{p}\rfloor$.

- 这里明显要分组，从每个组里抽效果要好。这里变的是A,B

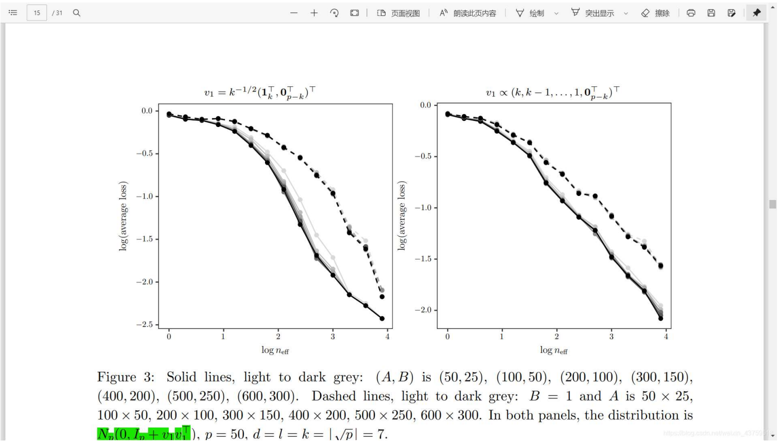

Figure 3: Solid lines, light to dark grey: (A, B) is $(50, 25)$, $(100, 50)$, $(200, 100)$, $(300, 150)$, $(400, 200)$, $(500, 250)$, $(600, 300)$. Dashed lines, light to dark grey: $B = 1$ and A is 50×25, 100×50, 200×100, 300×150, 400×200, 500×250, 600×300. In both panels, the distribution is $N_p(0, I_p + v_1 v_1^\top)$, $p = 50$, $d = l = k = \lfloor\sqrt{p}\rfloor = 7$.

在A,B里，固定一个变动另一个

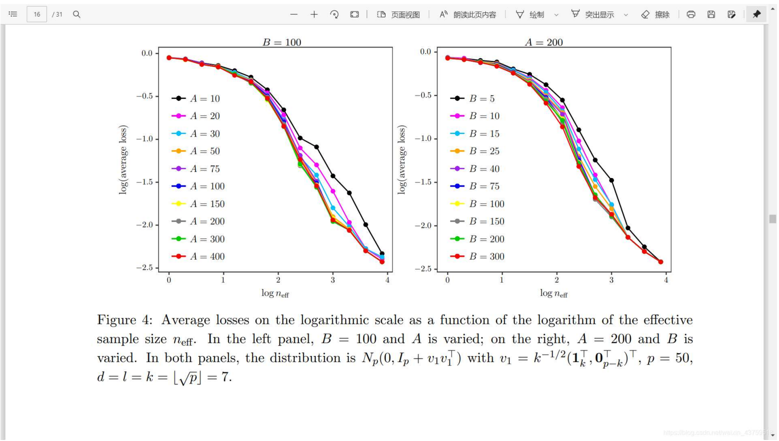

Figure 4: Average losses on the logarithmic scale as a function of the logarithm of the effective sample size n_{eff}. In the left panel, $B = 100$ and A is varied; on the right, $A = 200$ and B is varied. In both panels, the distribution is $N_p(0, I_p + v_1 v_1^\top)$ with $v_1 = k^{-1/2}(\mathbf{1}_k^\top, \mathbf{0}_{p-k}^\top)^\top$, $p = 50$, $d = l = k = \lfloor\sqrt{p}\rfloor = 7$.

对B的选择

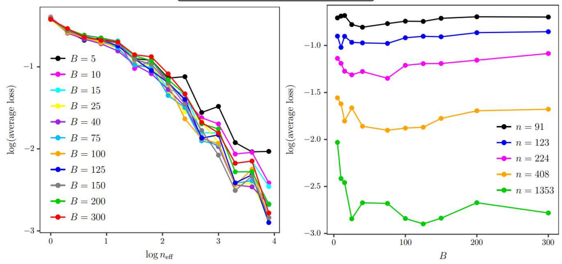

Figure 5: Trade-off in the choice of B. Left panel: the logarithm of average losses as a function of $\log n_{\text{eff}}$, where B is varied. Right panel: the logarithm of average losses as a function of B, where n is varied. In both panels, $A = 200$ and the distribution is $N_p(0, I_p + 5v_1v_1^\top + 4v_2v_2^\top)$ with $v_1 = k^{-1/2}(\mathbf{1}_k^\top, \mathbf{0}_{p-k}^\top)^\top$, $v_2 = k^{-1/2}(\mathbf{1}_3^\top, -1, 1, -1, 1, -1, 1, 1, \mathbf{0}_{p-k-3}^\top)^\top$, $p = 50$, $d = l = k = \lfloor\sqrt{p}\rfloor = 7$.

that A and B should increase with p. We suggest using $A = 300$ and $B = 100$ when $p \approx 100$, while $A = 600$ and $B = 200$ when $p \approx 1000$.

2. d,l

So far, in all our simulations, we have assumed that the true sparsity level k is known and we took $d = \ell = k$, where d is the dimension of the random projections and ℓ is the sparsity of the computed estimator. However, in practice k may not be known in advance. In Figure 6, we demonstrate how the over- and under-estimation of k affects the loss of our estimator. In

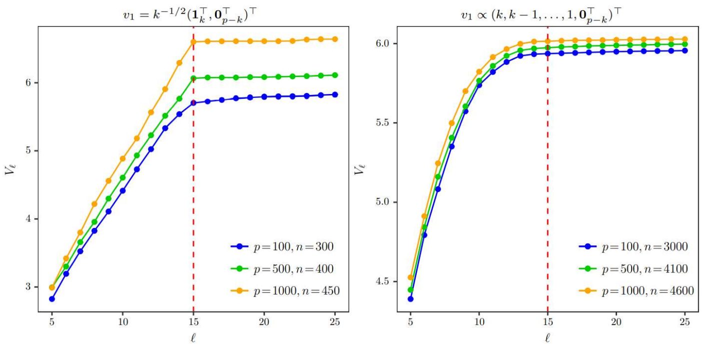

Figure 8: Selecting ℓ. Left: $\theta_1 = 5$, $\theta_2 = 3$, $k = 15$, $d = 10$, $A = 300$, $B = 100$. Right: $\theta_1 = 5$, $k = 15$, $d = 20$, $A = 500$, $B = 100$.

1.6.2. 与他人模型比较

- 第一主成分

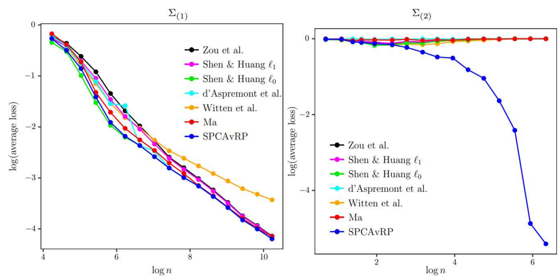

Figure 9: Loss against sample size n using two different covariance structures from (13). Blue: SPCAvRP with $A = 300, B = 150, d = l = k$; black: Zou, Hastie and Tibshirani (2006) with given k; magenta and green: Shen and Huang (2008) with ℓ_1 and ℓ_0-thresholding, respectively, both with given k; cyan: d'Aspremont, Bach and El Ghaoui (2008) with given k; orange: Witten, Tibshirani and Hastie (2009) with parameters chosen by their cross-validation; red: Ma (2013) with the default parameters.

这里厚道了些，把他人的参数用交叉验

- 高阶主成分

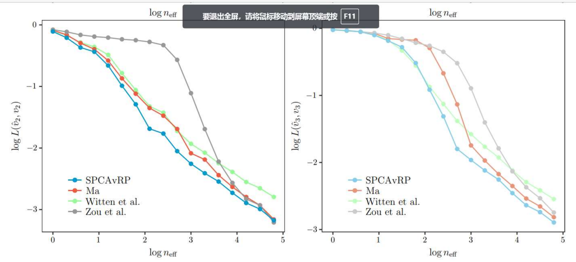

Figure 10: Observations are generated from $N_p(0,\Sigma)$, $\Sigma = I_p + \theta_1 v_1 v_1^\top + \theta_2 v_2 v_2^\top + \theta_3 v_3 v_3^\top$, $\theta_1 = 3$, $\theta_2 = 2$, $\theta_3 = 1$, $p = 100$, where v_1, v_2, v_3 have homogeneous signals strengths with $S_1 = \{1,\dots,10\}$, $S_2 = \{5,\dots,14\}$ and $S_3 = \{9,\dots,18\}$. The four figures correspond to loss functions $\|P_{\hat{V}_3} - P_{V_3}\|_{\text{op}}$, $L(\hat{v}_1, v_1)$, $L(\hat{v}_2, v_2)$ and $L(\hat{v}_3, v_3)$. SPCAvRP algorithm given in Algorithm 3 with input $A = 400$, $B = 200$, $s = 3$, $d = \ell_1 = \ell_2 = \ell_3 = k$, is compared with algorithms for subspace estimation proposed by Zou, Hastie and Tibshirani (2006), Witten, Tibshirani and Hastie (2009) and Ma (2013), which are used with their default parameters.

$S_1 \cap S_2 = \emptyset$	$\|P_{\hat{V}_2} - P_{V_2}\|_{\text{op}}$	$L(\hat{v}_1, v_1)$	$L(\hat{v}_2, v_2)$	$\|\hat{v}_1^\top \hat{v}_2\|$
SPCAvRP	5.69×10^{-2}	4.50×10^{-2}	5.49×10^{-2}	$< 10^{-15}$
Ma	8.47×10^{-2}	1.30×10^{-1}	1.37×10^{-1}	$< 10^{-15}$
Witten et al.	8.13×10^{-2}	9.75×10^{-2}	9.60×10^{-2}	4.61×10^{-3}
Zou et al.	9.60×10^{-2}	7.11×10^{-2}	9.39×10^{-2}	$< 10^{-15}$
$S_1 \cap S_2 \neq \emptyset$	$\|P_{\hat{V}_2} - P_{V_2}\|_{\text{op}}$	$L(\hat{v}_1, v_1)$	$L(\hat{v}_2, v_2)$	$\|\hat{v}_1^\top \hat{v}_2\|$
SPCAvRP	7.64×10^{-2}	7.97×10^{-2}	8.39×10^{-2}	$< 10^{-15}$
Ma	7.85×10^{-2}	1.31×10^{-1}	1.42×10^{-1}	$< 10^{-15}$
Witten et al.	9.20×10^{-2}	1.31×10^{-1}	1.33×10^{-1}	9.04×10^{-4}
Zou et al.	1.63×10^{-1}	1.84×10^{-1}	2.00×10^{-1}	5.84×10^{-4}

Table 1: Observations are generated from $N_p(0,\Sigma)$, $\Sigma = I_p + \theta_1 v_1 v_1^\top + \theta_2 v_2 v_2^\top$, $\theta_1 = 50$, $\theta_2 = 30$, $p = 200$, $n = 150$, where v_1 and v_2 have homogeneous signal strengths with $S_1 = \{1,\dots,14\}$, and $S_2 = \{15,\dots,28\}$ (top), $S_2 = \{7,\dots,20\}$ (bottom). The SPCAvRP algorithm given in Algorithm 3 with input $A = 300$, $B = 150$, $s = 2$, $d = \ell_1 = \ell_2 = k$, is compared with algorithms proposed by Zou, Hastie and Tibshirani (2006), Witten, Tibshirani and Hastie (2009) and Ma (2013), which are used with their default parameters.

1.6.3. Microarray数据

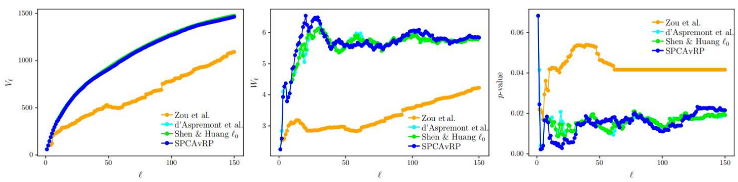

Figure 11: Left panel: V_ℓ; middle panel: Wasserstein distance W_ℓ between the empirical distributions of the two classes projected along $\hat{v}_{1,\ell}$; right panel: p-value of Welch's t-test for the two classes projected along $\hat{v}_{1,\ell}$, where $\hat{v}_{1,\ell}$ is the estimator of v_1 for varied sparsity level ℓ. For estimation we use SPCAvRP ($d = 30$, $A = 1200$, $B = 200$), Zou, Hastie and Tibshirani (2006), d'Aspremont, Bach and El Ghaoui (2008) and Shen and Huang (2008) with ℓ_0-thresholding.

这是箱型图

Shian150629

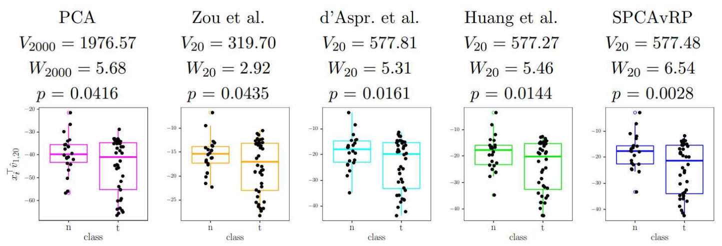

Figure 12: Variance V_ℓ, Wasserstein distance W_ℓ, p-value of the Welch's t-test and the corresponding box plots of the observations from the two classes projected along estimator $\hat{v}_{1,\ell}$ of the first principal component computed by five different approaches: classical PCA, Zou, Hastie and Tibshirani (2006), d'Aspremont, Bach and El Ghaoui (2008), Shen and Huang (2008) with ℓ_0-thresholding, and SPCAvRP . The desired sparsity level in all SPCA algorithms is set to $\ell = 20$.

总的来说结果还不错。论文写得蛮清楚，挺厚道的

2. 论文精读

说是精读但其实和RPENSEMBLE差不多，看代码为主

2.1. SPCA via random projections【第二章】

2.1.1. 单一主成分估计

2.1.2. 复合主成分估计

2.2. Numerical experiments【第四章】

3. 实验要点与思考

- loss function做一百次取均值
- SPCA前人的迭代算法，初值不好全完蛋