Homework 1

Yuan Yao

September 18, 2011

1. *Singular Value Decomposition:* The goal of this exercise is to refresh your memory about the singular value decomposition and matrix norms. A good reference to the singular value decomposition is Chapter 2 in this book:
   Matrix Computations, Golub and Van Loan, 3rd edition.
   Parts of the book are available online here:
   `http://books.google.com/books?id=ml0a7wPX6OYC&dq=Matrix+Computations&printsec=frontcover&source=bl&ots=lbfmg9JblY&sig=_ZNYb_4zdwfTrtn3zCKEZH9HVzA&hl=en&ei=cry8SuG7NMO2lAfUhtWYBA&sa=X&oi=book_result&ct=result&resnum=3#v=onepage&q=&f=false`

   (a) *Existence:* Prove the existence of the singular value decomposition. That is, show that if A is an $m \times n$ real valued matrix, then $A = U\Sigma V^T$, where U is $m \times m$ orthogonal matrix, V is $n \times n$ orthogonal matrix, and $\Sigma = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_p)$ (where $p = \min\{m, n\}$) is an $m \times n$ diagonal matrix. It is customary to order the singular values in decreasing order: $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_p \geq 0$. Determine to what extent the SVD is unique. (See Theorem 2.5.2, page 70 in Golub and Van Loan).

   (b) *Best rank-k approximation - operator norm:* Prove that the "best" rank-k approximation of a matrix in the operator norm sense is given by its SVD. That is, if $A = U\Sigma V^T$ is the SVD of A, then $A_k = U\Sigma_k V^T$ (where $\Sigma_k = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_k, 0, \ldots, 0)$ is a diagonal matrix containing the largest k singular values) is a rank-k matrix that satisfies

   $$\|A - A_k\| = \min_{\mathrm{rank}(B)=k} \|A - B\|.$$

   (Recall that the operator norm of A is $\|A\| = \max_{\|x\|=1} \|Ax\|$. See Theorem 2.5.3 (page 72) in Golub and Van Loan).

(c) *Best rank-k approximation - Frobenius norm:* Show that the SVD also provides the best rank-k approximation for the Frobenius norm, that is, $A_k = U\Sigma_k V^T$ satisfies

$$\|A - A_k\|_F = \min_{\text{rank}(B)=k} \|A - B\|_F.$$

(d) *Schatten p-norms:* A matrix norm $\|\cdot\|$ that satisfies

$$\|QAZ\| = \|A\|,$$

for all Q and Z orthogonal matrices is called a unitarily invariant norm. The Schatten p-norm of a matrix A is given by the ℓ_p norm ($p \geq 1$) of its vector of singular values, namely,

$$\|A\|_p = \left(\sum_i \sigma_i^p \right)^{1/p}.$$

Show that the Schatten p-norm is unitarily invariant. Note that the case $p = 1$ is sometimes called the nuclear norm of the matrix, the case $p = 2$ is the Frobenius norm, and $p = \infty$ is the operator norm.

(e) *Best rank-k approximation for unitarily invariant norms:* Show that the SVD provides the best rank-k approximation for any unitarily invariant norm. See also 7.4.51 and 7.4.52 in:
Matrix Analysis, Horn and Johnson, Cambridge University Press, 1985.

(f) *Closest rotation:* Given a square $n \times n$ matrix A whose SVD is $A = U\Sigma V^T$, show that its closest (in the Frobenius norm) orthogonal matrix R (satisfying $RR^T = R^T R = I$) is given by $R = UV^T$. That is, show that

$$\|A - UV^T\|_F = \min_{RR^T = R^T R = I} \|A - R\|_F,$$

where $A = U\Sigma V^T$. In other words, R is obtained from the SVD of A by dropping the diagonal matrix Σ. Use this observation to conclude what is the optimal rotation that aligns two sets of points $p_1, p_2, \ldots, p_n$ and $q_1, \ldots, q_n$ in $\mathbb{R}^d$, that is, find R that minimizes $\sum_{i=1}^n \|Rp_i - q_i\|^2$. See also (the papers are posted on course website):

- [Arun87] Arun, K. S., Huang, T. S., and Blostein, S. D., "Least-squares fitting of two 3-D point sets", *IEEE Transactions on Pattern Analysis and Machine Intelligence,* **9** (5), pp. 698–700, 1987.

- [Keller75] Keller, J. B., "Closest Unitary, Orthogonal and Hermitian Operators to a Given Operator", *Mathematics Magazine*, **48** (4), pp. 192–197, 1975.

- [FanHoffman55] Fan, K. and Hoffman, A. J., "Some Metric Inequalities in the Space of Matrices", *Proceedings of the American Mathematical Society*, **6** (1), pp. 111–116, 1955.

2. *James-Stein Estimators:* Suppose that $X_i \sim \mathcal{N}(\mu, I_p)(i = 1 \dots n)$ are independent p-Gaussian variables, let $y = \hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n X_i \sim \mathcal{N}(\mu, \frac{1}{\sqrt{n}} I_p) = \mathcal{N}(\mu, \varepsilon I_p)$, $\varepsilon = \frac{1}{\sqrt{n}}$, consider the James-Stein estimator

$$\tilde{\mu}_{JS} = \left(1 - \frac{\varepsilon^2(p-2)}{\|\hat{\mu}_n\|^2}\right)\hat{\mu}_n$$

and its positive part truncation

$$\tilde{\mu}_{JS+} = \left(1 - \frac{\varepsilon^2(p-2)}{\|\hat{\mu}_n\|^2}\right)_+ \hat{\mu}_n,$$

where $(x)_+ = max(0, x)$.

(a) Show that

$$\mathbb{E}\|\tilde{\mu}_{JS} - \mu\|^2 \le \mathbb{E}\|\hat{\mu}_n - \mu\|^2 - \frac{\varepsilon^4(p-2)^2}{\varepsilon^2 p + \|\mu\|^2}$$

so when $p >> 2$ and $\mu = 0$, the gain will be in $O(\varepsilon^2 p)$. (Hint: use Jenson inequality as Lemma 3.10 in [Tsybakov09]).

(b) Show that

$$\mathbb{E}\|\tilde{\mu}_{JS+} - \mu\|^2 \le \mathbb{E}\|\tilde{\mu}_{JS} - \mu\|^2.$$

(Hint: see Lemma A.6 in Appendix of [Tsybakov09])

3. *Phase transition in PCA "spike" model:* Consider a finite sample of n i.i.d vectors $x_1, x_2, \dots, x_n$ drawn from the p-dimensional Gaussian distribution $\mathcal{N}(0, \sigma^2 I_{p\times p} + \lambda_0 uu^T)$, where λ_0/σ^2 is the signal-to-noise

ratio (SNR) and $u \in \mathbb{R}^p$. In class we showed that the largest eigenvalue λ of the sample covariance matrix S_n

$$S_n = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T$$

pops outside the support of the Marcenko-Pastur distribution if

$$\frac{\lambda_0}{\sigma^2} > \sqrt{\gamma},$$

or equivalently, if

$$\text{SNR} > \sqrt{\frac{p}{n}}.$$

(Notice that $\sqrt{\gamma} < (1+\sqrt{\gamma})^2$, that is, λ_0 can be "buried" well inside the support Marcenko-Pastur distribution and still the largest eigenvalue pops outside its support). All the following questions refer to the limit $n \to \infty$ and to almost surely values:

(a) Find λ given SNR $> \sqrt{\gamma}$.

(b) Use your previous answer to explain how the SNR can be estimated from the eigenvalues of the sample covariance matrix.

(c) Find the squared correlation between the eigenvector v of the sample covariance matrix (corresponding to the largest eigenvalue λ) and the "true" signal component u, as a function of the SNR, p and n. That is, find $|\langle u, v\rangle|^2$.

(d) Confirm your result using MATLAB simulations (e.g. set $u = e$; and choose $\sigma = 1$ and λ_0 in different levels. Compute the largest eigenvalue and its associated eigenvector, with a comparison to the true ones.)

4. *Finite rank perturbations of random symmetric matrices:* Wigner's semi-circle law (proved by Eugene Wigner in 1951) concerns the limiting distribution of the eigenvalues of random symmetric matrices. It states, for example, that the limiting eigenvalue distribution of $n \times n$ symmetric matrices whose entries w_{ij} on and above the diagonal ($i \le j$) are i.i.d Gaussians $\mathcal{N}(0, \frac{1}{4n})$ (and the entries below the diagonal are determined by symmetrization, i.e., $w_{ji} = w_{ij}$) is the semi-circle:

$$p(t) = \frac{2}{\pi}\sqrt{1-t^2}, \quad -1 \le t \le 1,$$

where the distribution is supported in the interval $[-1, 1]$.

(a) Confirm Wigner's semi-circle law using MATLAB simulations (take, e.g., $n = 400$).

(b) Find the largest eigenvalue of a rank-1 perturbation of a Wigner matrix. That is, find the largest eigenvalue of the matrix

$$W + \lambda_0 u u^T,$$

where W is an $n \times n$ random symmetric matrix as above, and u is some deterministic unit-norm vector. Determine the value of λ_0 for which a phase transition occurs. What is the correlation between the top eigenvector of $W + \lambda_0 u u^T$ and the vector u as a function of λ_0? Use techniques similar to the ones we used in class for analyzing finite rank perturbations of sample covariance matrices.

5. *PCA experiments:* Take any digit data ('0',...,'9'), or all of them, from website

`http://www-stat.stanford.edu/~tibs/ElemStatLearn/datasets/zip.digits/`

and perform PCA experiments with Matlab or other language you are familiar:

(a) Set up data matrix $X = (x_1, \ldots, x_n) \in \mathcal{R}^{p \times n}$;

(b) Compute the sample mean $\hat{\mu}_n$ and form $\tilde{X} = X - e\hat{\mu}_n^T$;

(c) Compute top k SVD of $\tilde{X} = U S_k V^T$;

(d) Plot eigenvalue curve, i.e. i vs. $\lambda_i(\hat{\Sigma}_n)/tr(\hat{\Sigma}_n)$ $(i = 1, \ldots, k)$, with top-k eigenvalue λ_i for sample covariance matrix $\hat{\Sigma}_n = \frac{1}{n}\tilde{X} * \tilde{X}^T$;

(e) Use `imshow` to visualize the mean and top-k principle components as *left* singular vectors $U = [u_1, \ldots, u_k]$;

(f) For $k = 1$, sort the image data (x_i) $(i = 1, \ldots, n)$ according to the top *right* singular vectors, v_1, in an ascending order;

(g) For $k = 2$, scatter plot (v_1, v_2) and select a grid on such a plane to show those images on the grid (e.g. Figure 14.23 in book [ESL]: Elements of Statistical Learning).