NOTES FOR 21 NOV (THURSDAY)

1. Recap

(1) Proved the spectral theorem for normal operators.
(2) Proved Sylvester's law of inertia for real quadratic forms.

2. Bilinear forms

Actually, we have a similar result for general vector spaces.

Theorem 2.1. *Let V be a finite-dimensional vector space over a field of characteristic $\neq 2$ and let T be a symmetric bilinear form. Then there is an ordered basis of V in which T is represented by a diagonal matrix. If the field is $\mathbb{C}$, the matrix can be chosen to consist only of 0s and 1s.*

Proof. Assume that $T \neq 0$ (if not, it is trivial) and let $n = dim(V)$. We induct on n. (For $n = 1$ it is trivial). Let e be a vector such that $T(e,e) \neq 0$ (if $\mathbb{F} = \mathbb{C}$, then redefine e to be $\frac{e}{\sqrt{T(e,e)}}$ so that its "T"-norm is 1). Define W to be the subspace of vectors such that $T(e,w) = 0$. We claim that $V =< e > \oplus W$. Indeed, their intersection is trivial. Now, $v = \frac{T(v,e)}{T(e,e)}e + (v - \frac{T(v,e)}{T(e,e)}e)$ which is in $< e > + W$. By induction (because T restricted to W is still a symmetric bilinear form), assume that there is a basis e_j of W such that T is already in the chosen form in that basis. Now $T(e,e_j) = 0$. Hence T is diagonal in the basis e, e_j. (In the complex case, the diagonal elements are 0 or 1.)
(Where did we use the assumption that $char(\mathbb{F}) \neq 2$?) □

More generally, if T is a bilinear form (not necessarily symmetric), its rank is defined as the rank of its matrix (in any basis) and likewise its nullity. This concept can be phrased in a basis-free manner. (The following simple concept (and versions of it) are important in algebraic topology (Poincaré duality) and differential geometry (symplectic form).)

Theorem 2.2. *Let T be a bilinear form on a finite-dimensional vector space V. Let $L_T, R_T : V \to V^*$ be the maps defined as $L_T v(w) = T(v,w)$ and $R_T v(w) = T(w,v)$. Then, L_T, R_T are linear maps and $rank(L_T) = rank(R_T) = rank(T)$. Moreover, TFAE.*

(1) *$rank(T) = dim(V)$.*
(2) *For each non-zero $v \in V$ there exists a $w \in V$ such that $f(v,w) \neq 0$ and vice-versa. (This is called the property of non-degeneracy.)*

Proof. Obviously L_T, R_T are linear. The kernel of L_T consists of v such that $T(v,w) = 0 \ \forall \ w \in V$, i.e., $w^T B v = 0$ and hence, $Bv = 0$ which is simply the nullity of T. By nullity-rank, $Rank(L_T) = Rank(B) = Rank(T) = Rank(B^T) = Rank(R_T)$.
If T has full rank, then kernel of L_T is trivial. Hence, for every non-zero $v \in V$, there is a w such that $T(v,w) = L_T v(w) \neq 0$. Likewise for R_T. If non-degeneracy holds, then clearly the kernel of L_T is trivial and T has full rank. □

Recall that Hermitian matrix (or alternatively, a Hermitian sesquilinear form) A is said to be positive-definite if $v^\dagger A v > 0$ for all $v \neq 0$. Here is a theorem that gives a criterion to check for positivity.

Theorem 2.3. *A Hermitian $n \times n$ matrix A is positive-definite iff its principal minors, $\Delta_k(A) = \det(A_{ij})$ where $1 \leq i \leq k$ are all positive.*

Proof. If the principal minors are positive : Assume inductively that the statement is true for $1, 2 \ldots, n-1$ (for n=1 it is trivial). Since positive-definiteness is equivalent to the eigenvalues being positive, diagonalise (using a unitary matrix that keeps the last basis vector of $\mathbb{C}^n$ intact) the principal $n-1 \times n-1$ part of the matrix to see that the eigenvalues be $\lambda_1, \ldots, \lambda_{n-1}$ (Note that these are NOT necessarily eigenvalues of the bigger matrix) are all positive by the induction assumption. That is $UAU^\dagger = \tilde{A}$. Note that A is positive-definite iff $\tilde{A}$ is so. (Why ?) Also, $\det(A) = \det(\tilde{A})$. (Why ?)

The determinant of the bigger matrix can be easily computed to be equal to $\lambda_1 \lambda_2 \ldots \lambda_{n-1}(A_{nn} - \sum_i \frac{|A_{ni}|^2}{\lambda_i})$ which is positive by assumption. Let v be a vector written in the new basis (the eigenvectors of the principal $n-1 \times n-1$ matrix and e_n). Then $v^\dagger A v = |v_n|^2 A_{nn} + \sum_k A_{nk} v_k \bar{v}_n + \sum_k \bar{A}_{nk} \bar{v}_k v_n + \sum_i \lambda_i |v_i|^2$ which is positive for all non-zero v_n if the discriminant is negative. The discriminat is proportional to $|\sum_k A_{nk} v_k|^2 - A_{nn} \sum_i \lambda_i |v_i|^2 < 0$ by Cauchy-Schwarz on the first term and $A_{nn} - \sum_i \frac{|A_{ni}|^2}{\lambda_i} > 0$.

If A is positive definite, then so is $UAU^\dagger = \tilde{A}$ (after diagonalising). Hence, so is the principal $n-1 \times n-1$ part of $\tilde{A}$. Hence, the eigenvalues of the principal $n-1 \times n-1$ part of $\tilde{A}$, i.e., $\lambda_1 = \tilde{A}_{11}, \ldots, \lambda_{n-1} = \tilde{A}_{n-1 n-1}$, which are the eigenvalues of the principal $n-1 \times n-1$ part of A are all positive. Inductively, this fact means that the principal k minors where $k \neq n-1$ of A are positive. Moreover, since $\tilde{A}$ is positive-definite, its eigenvalues are all positive and hence its determinant (which is the product of its eigenvalues) are positive. Hence, all the principal minors of A are positive. □

3. Applications

3.1. **Least squares.** Plotting the prices of houses vs their surface area in a given locality, we expect the data to lie on a straight line, but often this is not exactly the case. So what is the best-fitting straight line ? In general, suppose we expect a dependent variable $y \in \mathbb{R}$ to depend linearly on some independent variables x_i as $y = \sum_i a_i x_i + a_{n+1}$, then firstly, by introducing another independent constant "variable" $x_{n+1} = 1$, $y = \sum_i a_i x_i$. Each observation can lead to $y_j = \sum_i a_i x_{ij} + \epsilon_j$ where ϵ_j is a random error. We can write this equation as $\vec{y} = X\vec{a} + \vec{\epsilon}$ where $\vec{a}$ is the vector of "weights"/coefficients and the rows of the matrix X consist of data for each observation. A natural thing to do is to choose the weights so that the squared error $S = \|\vec{y} - X\vec{a}\|^2$ is minimum. There are two ways to approach this problem - calculus and linear algebra. In the second approach, we want to find the best approximation to $\vec{y}$ from the subspace formed by the columns of X. That vector is the orthogonal projection onto the column space of X, i.e., $(y - Xa)^T Xb = 0$ for all b, i.e., $y^T X = a^T X^T X$ or $X^T y = X^T X a$ (the so-called normal equations). Note that $X^T X$ is positive-semidefinite. If it is actually positive-definite, it is invertible and $a = (X^T X)^{-1} X^T y$. It can fail to be invertible when there is too little data and too many parameters (overfitting). In such cases, one can add $X^T X + cI$ where $c > 0$ is a small positive number to make it invertible. (Corresponds to minimising $\|\vec{y} - X\vec{a}\|^2 + c a^t a$.) This method is called Tikhonov regularisation in Machine Learning. The parameter c is chosen to be small enough so that the solution is not too bad, but large enough so that overfitting does not occur.

Another reason for minimising the least square error comes from statistics : Assuming that ϵ are iid Gaussian normal variables with 0 mean and variance σ^2, the ln of the probability density for $\vec{\epsilon}$ to be $\vec{y} - X\vec{a}$ is proportional to $-\|y - X\vec{a}\|^2$ which is maximised precisely when the normal equations hold.

3.2. **Principal Component Analysis.** It is hard to store too much data. Suppose we are given 500 photos of the same person giving different expressions, is there a way to store only a little important information to approximately reconstruct the 500 photos ? (Clearly these sorts of problems can have ramifications in automatic tagging on Facebook.) That is this information must capture most of the variation in the 500 photos. We want a linear map that transforms the data (given as matrix X with column-wise zero average, each row vector $\vec{x}_i$ corresponds to a different photo/different repetition of an experiment, and each column to a feature (i.e. pixel intensity or colour or in the case of experiments, to values from different sensors for instance)) to a new set of l-dimensional vectors $\vec{t}_i$ (called the PCA scores) using p-dimensional unit weight vectors $\vec{w}_k$ such that $t_{k(i)} = \langle \vec{x}_i, \vec{w}_k \rangle$ where $i = 1, \ldots, n$ and $k = 1, \ldots, l$, such that the individual new features $t_1, \ldots, t_l$ inherit most of the variance from X.

The first weight vector $\vec{w}_1$ should maximise the variance inherited to t_1, i.e., $\max_{||w||=1} \sum_i (t_1)^2_{(i)} = \max_{||w||=1} \sum_i \langle \vec{x}_i, \vec{w} \rangle^2 = \max_{||w||=1} w^T X^T X w$ which is simply the largest eigenvector of $X^T X$. Likewise, it makes sense for the other weight vectors to be the other eigenvectors of $X^T X$. The PCA scores are $T = XW$. Typically, one keeps only a few singular vectors (and hence reconstruction is not perfect). Clearly, the SVD plays a role here.

Here is an actual example where this procedure is implemented in a different manner (In what I outlined above, you store 500 images but not all pixels, in this url, they store all pixels but fewer images that they call eigenimages) : http://people.ciirc.cvut.cz/~hlavac/TeachPresEn/11ImageProc/15PCA.pdf.