

Homework 4, Stat 541: Due Friday, Oct 16, 2009, 12 noon

Linear Algebra and Linear Models

Student Name: (replace this with your name)

October 6, 2009

Instructions: Edit this LaTeX file with your solutions and generate a PDF file from it. E-mail the PDF to the usual class gmail address.

For some questions one needs to understand basis changes and associated coordinate transformations. To brush up (or to finally really understand what that is), you may want to check the solutions of Homework 2, Problem 11. See also Strang's text on linear algebra, the Appendix "Linear Transformations, Matrices, and Change of Basis."

The usual honor code applies (see the class webpage). In particular, it is strictly prohibited to consult previous years' solutions.

1. Interpretation of eigendecompositions: You know that for any real symmetric $p \times p$ matrix \mathbf{S} there exists an orthonormal basis $(\mathbf{u}_j)_{j=1\dots p}$ of eigenvectors and associated eigenvalues $(\lambda_j)_{j=1\dots p}$ which can be assumed in descending order w.l.o.g.: $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$. State what it means for \mathbf{u}_j to be an eigenvector with eigenvalue λ_j , both formally (state the definition) and geometrically (how does an eigenvector move under \mathbf{S} ?). Comment on what it means if $\lambda_j = 1$ or $\lambda_j = -1$ or $\lambda_j = 0$.

Answer: The definition is $\mathbf{S}\mathbf{u}_j = \lambda_j\mathbf{u}_j$. Geometrically this means the linear transformation given by \mathbf{S} only moves an eigenvector in its own direction, $\mathbf{u}_j \mapsto \lambda_j\mathbf{u}_j$, but doesn't move it "sideways."

- If $\lambda_j = 1$ it means that \mathbf{u}_j is a fixed point under \mathbf{S} : $\mathbf{S}\mathbf{u}_j = \mathbf{u}_j$.

- If $\lambda_j = 0$ it means that \mathbf{u}_j is in the null space of \mathbf{S} : $\mathbf{S}\mathbf{u}_j = \mathbf{0}$.
 - If $\lambda_j = -1$ it means that \mathbf{u}_j gets ‘flipped’ at the origin: $\mathbf{S}\mathbf{u}_j = -\mathbf{u}_j$.
2. Under the assumption that the vectors \mathbf{u}_j form an orthonormal system of eigenvectors of \mathbf{S} , form the matrix $\mathbf{U} = (\mathbf{u}_1, \dots, \mathbf{u}_p)$ and form the diagonal matrix Λ that has the eigenvalues in the diagonal (zero off-diagonal). Express what orthonormality of the vectors \mathbf{u}_j means in terms of \mathbf{U} as a matrix equation. Then express the p eigenvalue conditions with a single matrix equation in terms of \mathbf{S} , \mathbf{U} and Λ . (This is standard textbook material.)

Answer: Orthonormality of the vectors \mathbf{u}_j means that $\mathbf{u}_j^T \mathbf{u}_k = 0$ ($j \neq k$) and $\mathbf{u}_j^T \mathbf{u}_j = 1$, but these values are the (j, k) off-diagonal and j 'th diagonal element of $\mathbf{U}^T \mathbf{U}$. Hence orthonormality can be expressed as $\mathbf{U}^T \mathbf{U} = \mathbf{I}$, which just means that \mathbf{U} is an orthogonal matrix. The p eigenvalue equations can be expressed in a single matrix equation as follows: $\mathbf{S}\mathbf{U} = \mathbf{U}\Lambda$.

3. Again under the assumption that the vectors \mathbf{u}_j form an orthonormal system of eigenvectors of \mathbf{S} , form the matrices $\mathbf{P}_j = \mathbf{u}_j \mathbf{u}_j^T$ and explain what type of linear map they describe (refer to HW2). Then express \mathbf{S} in terms of $\mathbf{P}_1, \dots, \mathbf{P}_p$ and $\lambda_1, \dots, \lambda_p$. (Hint: \mathbf{S} is just a linear combination of the maps \mathbf{P}_j .)

Answer: The matrix \mathbf{P}_j represents the orthogonal projection onto the direction of the eigenvector \mathbf{u}_j . The matrix eigenequation of the previous problem can be written as $\mathbf{S} = \mathbf{U}\Lambda\mathbf{U}^T$ because $\mathbf{U}^T = \mathbf{U}^{-1}$ due to orthogonality of \mathbf{U} . Similar to Problem 1(a) of Homework 2, we see that the right hand side can be re-expressed as $\mathbf{S} = \sum_j \mathbf{u}_j \lambda_j \mathbf{u}_j^T$ which is the same as $\mathbf{S} = \sum_j \lambda_j \mathbf{P}_j$. Hence a symmetric matrix is the “eigenvalue-weighted sum” of orthogonal projections onto the eigenvectors.

4. What is the name of the property of \mathbf{S} called that is equivalent to $\lambda_p \geq 0$? Same for $\lambda_p > 0$? (This is just regurgitating textbook material.)

Answer: The former is “non-negative definiteness,” the latter “positive definiteness.”

5. What is the matrix of the linear transformation given by \mathbf{S} in the new coordinate system after changing the basis to $\mathbf{u}_1, \dots, \mathbf{u}_p$?

Answer: It is

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & 0 & \dots & 0 \\ 0 & \lambda_2 & 0 & \dots & 0 \\ 0 & 0 & \lambda_2 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & \lambda_p \end{pmatrix}$$

The complex reasoning is to go through the mechanics of a basis change and ensuing coordinate transformation as in Homework 2, Problem 11, using $\mathbf{A} = \mathbf{U}^T$. A simpler reasoning is to express $\mathbf{S}\mathbf{u}_j$ in the basis $\mathbf{u}_1, \dots, \mathbf{u}_p$. The resulting coefficients are the coordinates of the former in the latter and form the columns of the linear transformation in the new basis. A higher triviality results: $\mathbf{S}\mathbf{u}_j = \lambda_j\mathbf{u}_j$ means the j 'th column of the new matrix has all zeros except λ_j in the j 'th entry.

6. The following is called the Rayleigh quotient of \mathbf{S} :

$$R(\mathbf{u}) = \frac{\mathbf{u}^T \mathbf{S} \mathbf{u}}{\|\mathbf{u}\|^2}$$

Find and interpret the stationary equations for $R(\mathbf{u})$: $\partial_{\mathbf{u}} R(\mathbf{u}) = \mathbf{0}$ ($\partial_{\mathbf{u}}$ = gradient w.r.t. \mathbf{u}).

Answer: The gradient is

$$\partial_{\mathbf{u}} R(\mathbf{u}) = \frac{2\mathbf{S}\mathbf{u}(\|\mathbf{u}\|^2) - (\mathbf{u}^T \mathbf{S} \mathbf{u}) 2\mathbf{u}}{\|\mathbf{u}\|^4},$$

and forcing it to zero means the numerator is zero:

$$2\mathbf{S}\mathbf{u}(\|\mathbf{u}\|^2) - (\mathbf{u}^T \mathbf{S} \mathbf{u}) 2\mathbf{u} = \mathbf{0}.$$

Canceling the factors 2 and dividing by $\|\mathbf{u}\|^2$ gives us the following pretty form:

$$\mathbf{S}\mathbf{u} = R(\mathbf{u}) \mathbf{u}.$$

This can be interpreted as the matrix eigenvalue equation with $\lambda = R(\mathbf{u})$. As a consequence, the eigenvectors of \mathbf{S} are exactly the stationary directions of the Rayleigh quotient. From this we get an insight: the eigenvalue of an eigenvector equals the Rayleigh quotient of the eigenvector.

7. Assume \mathbf{u}_1 and \mathbf{u}_2 are orthonormal eigenvectors of \mathbf{S} with identical eigenvalues $\lambda_1 = \lambda_2$. Consider the following one-parameter family of linear combinations $\mathbf{u}(t) = \cos(t)\mathbf{u}_1 + \sin(t)\mathbf{u}_2$. What geometric shape is traced by $\mathbf{u}(t)$? What can you say about $\mathbf{u}(t)$ in relation to \mathbf{S} ? In what relation is $\mathbf{u}(t)$ relative to the eigenvectors $\mathbf{u}_3, \dots, \mathbf{u}_p$?

Answer: The vectors $\mathbf{u}(t) = \cos(t)\mathbf{u}_1 + \sin(t)\mathbf{u}_2$ form a unit circle in the 2-dimensional subspace (“plane”) spanned by \mathbf{u}_1 and \mathbf{u}_2 : $\|\mathbf{u}(t)\| = 1$ for all t (check!). All vectors $\mathbf{u}(t)$ are also eigenvectors of \mathbf{S} for the common eigenvalue $\lambda_1 = \lambda_2$. We also have $\langle \mathbf{u}(t), \mathbf{u}_j \rangle = 0$ for all $j > 2$, hence we have a whole circle’s worth of eigenvectors orthogonal to the remaining eigenvectors.

8. What is the eigendecomposition of an orthogonal projection of rank r ? How unique is it?

Answer: The eigenvalues can only be 1 and 0 for an idempotent linear map because $P^2 = P$ implies $\lambda^2 = \lambda$. For finding the eigenvectors it is natural to look to the range space and null space that uniquely characterize an idempotent map: $\mathcal{R}(\mathbf{P}) = \{ \mathbf{P}\mathbf{v} \mid \mathbf{v} \in \mathbb{R}^p \}$ and $\mathcal{N}(\mathbf{P}) = \{ \mathbf{v} \mid \mathbf{P}\mathbf{v} = \mathbf{0} \}$, which for orthogonal projections are orthogonal to each other. We have: $\mathbf{v} \in \mathcal{R}(\mathbf{P})$ satisfy $\mathbf{P}\mathbf{v} = \mathbf{v}$, hence are eigenvectors with eigenvalue $\lambda = 1$; $\mathbf{v} \in \mathcal{N}(\mathbf{P})$ satisfy $\mathbf{P}\mathbf{v} = \mathbf{0}$, hence are eigenvectors with eigenvalue $\lambda = 0$. It is therefore sufficient to pick an orthonormal basis of $\mathcal{R}(\mathbf{P})$ and an orthonormal basis of $\mathcal{N}(\mathbf{P})$ which together form an orthonormal basis of \mathbb{R}^p , and the eigenvalues are 1 for the $\mathcal{R}(\mathbf{P})$ basis and 0 for the $\mathcal{N}(\mathbf{P})$ basis.

9. What is the matrix of an orthogonal projection after a change to a basis of eigenvectors? (This question is almost redundant with the previous problem.)

Answer: We know from a previous problem for the general case of a symmetric matrix \mathbf{S} that the new matrix is Λ . In this particular case the diagonal matrix has 1’s in the first

Answer: $\mathbf{P} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$, hence $P_{ii} = \sum_j U_{ji}\lambda_j U_{ji} = \sum_{j \leq \text{rank}(\mathbf{P})} U_{ji}^2$ because $\lambda_1 = \dots = \lambda_{\text{rank}(\mathbf{P})} = 1$ and the rest is zero. We know that $\sum_j U_{ij}^2 = 1$, hence P_{ii} must be between 0 and 1.

Here is an even better proof that only uses idempotence and not symmetry, hence holds even for non-orthogonal projections: From $P^2 = P$ follows

$$P_{ii} = \sum_j P_{ij}^2 \geq P_{ii}^2.$$

Now the inequality $x \geq x^2$ is equivalent to $0 \leq x \leq 1$. QED.

15. Consider a simple linear regression where we subtracted the mean from the response vector and from the predictor vector.
- Write down the projection matrix \mathbf{P} onto the centered predictor vector $\mathbf{x} \in \mathbb{R}^N$.
 - Comment on the “total budget” in the diagonal of \mathbf{P} .
 - Interpret the diagonal elements in terms of self-influence.
 - Interpret the diagonal elements in terms of variance of fitted values and residuals.
 - Interpret the off-diagonal elements in terms of correlation between pairs of fitted values and pairs of residuals.

Answer:

- The projection matrix is $\mathbf{P} = \mathbf{x}\mathbf{x}^T / \|\mathbf{x}\|^2$.
- The “total budget” in the diagonal of \mathbf{P} is 1 because this is a one-dimensional projection corresponding to a single degree of freedom for the fitted slope that belongs to the predictor \mathbf{x} . While we already know this from the general case, one can immediately see the “total budget” from this particular trace: $\sum_i P_{ii} = \sum_i x_i^2 / \|\mathbf{x}\|^2 = 1$.
- The i 'th diagonal element is $P_{ii} = x_i^2 / \|\mathbf{x}\|^2$. It measures the squared distance of the i 'th case from the origin in predictor space (here just the real line), that is, from the mean (since \mathbf{x} is centered). Hence cases for which x_i is far from the mean have more self-influence than cases near the mean. Self-influence increases with the square of the distance from the mean, which is rather rapidly.

(d) Ignoring a common factor σ^2 , the variance of a fitted value \hat{y}_i is P_{ii} , and that of a residual r_i is $1 - P_{ii}$. The variance of a fitted value is the larger the farther the case is from the mean of \mathbf{x} , and conversely the variance of a residual is the smaller the farther the case is from the mean of \mathbf{x} . Again, this reflects the effects of self-influence: cases that are distant in predictor space are more self-influential and hence tend to have smaller residuals, while their fitted values tend to bounce around more across datasets.

(e) The off-diagonal elements are $P_{ij} = x_i x_j / \|\mathbf{x}\|^2$. They are positive if x_i and x_j are on the same side of the origin (= the mean), and negative if on opposite sides. Accordingly, the correlation between \hat{y}_i and \hat{y}_j is positive or negative, and the correlation between r_i and r_j is negative or positive, respectively. This illustrates the “seesaw effect” of fitted lines with fulcrum at the mean of the predictor variable.

16. If $V[X]$ is the covariance matrix of a p -dimensional random vector X , express $\text{Var}[\mathbf{a}^T X]$ in terms of $V[X]$ and \mathbf{a} , and assume that \mathbf{a} is a unit vector. If you know $V[X]$, what do you know about the distribution of X in all directions?

Answer: $\text{Var}[\mathbf{a}^T X] = \mathbf{a}^T V[X] \mathbf{a}$, where we can think of $\mathbf{a}^T X$ as the projection of X onto the direction \mathbf{a} . Hence if we know $V[X]$, we know how much X is spread out in any direction \mathbf{a} as measured by variance.

17. Two non-negative matrices \mathbf{S}_1 and \mathbf{S}_2 are said to be ordered, “ $\mathbf{S}_1 \geq \mathbf{S}_2$ ” if $\mathbf{S}_1 - \mathbf{S}_2$ is non-negative definite. If we are given two p -dimensional random vectors X and Y , and if $V[X] \geq V[Y]$, what does this mean in view of the previous problem?

Answer: It means that X is more spread out than Y in all directions \mathbf{a} as measured by variance: $\text{Var}[\mathbf{a}^T X] - \text{Var}[\mathbf{a}^T Y] = \mathbf{a}^T V[X] \mathbf{a} - \mathbf{a}^T V[Y] \mathbf{a} = \mathbf{a}^T (V[X] - V[Y]) \mathbf{a} \geq 0$.

18. In a linear regression problem let $\mathbf{x}_1, \dots, \mathbf{x}_N$ be the “feature vectors”, that is, the rows of the predictor matrix written as $(p + 1)$ -dimensional column vectors. Assume that they are N i.i.d. samples from the $(p + 1)$ -dimensional random vector X (and you can even allow the first coordinate to be all ones for the intercept, thought of as i.i.d. samples from the pointmass at 1.0). Q: Is it desirable that $V[X]$ be “large” or “small” in the sense of

the ordering of symmetric matrices defined in the previous problems? If it helps use the notation $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_N)^T$ for the $N \times (p+1)$ predictor matrix. (No need to prove anything about ordering of inverse matrices; you can argue that point intuitively.)

Answer: Large. The reason can be seen from the standard error variance formula for the coefficient vector: $V[\mathbf{b}] = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} \approx \sigma^2V[X]^{-1}/N$ (or $\dots/(N-1)$, which is asymptotically indistinguishable). We want the regression coefficients $\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ to be precise, hence $\text{Var}[\mathbf{a}^T\mathbf{b}]$ to be small for all fixed \mathbf{a} , hence $V[\mathbf{b}]$ small in the sense of the previous problems, hence $V[X]^{-1}$ small, hence $V[X]$ large (this is the “intuitive” part).

19. What does adjustment of a $N \times p$ data matrix \mathbf{X} with regard to $(1, \dots, 1)^T \in \mathbb{R}^N$ mean?

Answer: Adjusting for this vector means subtracting the mean: the projection \mathbf{P} onto this vector is a square matrix with all 1's, divided by n :

$$\mathbf{P} = \frac{1}{\|\mathbf{e}\|^2} \mathbf{e}\mathbf{e}^T = \frac{1}{n} \begin{pmatrix} 1 & 1 & \dots \\ 1 & 1 & \dots \\ \dots & \dots & \dots \end{pmatrix}$$

where $\mathbf{e}^T = (1, 1, \dots, 1) \in \mathbb{R}^N$. If we denote by \mathbf{x}_j the j 'th column of \mathbf{X} , then $\mathbf{P}\mathbf{x}_j = (\text{mean}(\mathbf{x}_j), \dots, \text{mean}(\mathbf{x}_j))^T \in \mathbb{R}^N$. Subtracting this from \mathbf{x}_j means centering \mathbf{x}_j .

20. Write down the projection matrix for a categorical variable with three categories and category sizes $N_1 + N_2 + N_3 = N$. Assume that the cases are listed category $k = 1$ first, then $k = 2$, then $k = 3$, so that you can use a block-diagonal notation with blocks of size $N_k \times N_k$. Finally, indicate what the matrix is that adjusts for this categorical variable.

Answer: This is the same as the previous problem except in three blocks: We need to form category means which are obtained by projection as follows: Let \mathbf{E}_k be a $N_k \times N_k$ matrix filled with 1's. Then the projection is

$$\mathbf{P} = \begin{pmatrix} \frac{1}{N_1}\mathbf{E}_1 & 0 & 0 \\ 0 & \frac{1}{N_2}\mathbf{E}_2 & 0 \\ 0 & 0 & \frac{1}{N_3}\mathbf{E}_3 \end{pmatrix}$$

Projecting with \mathbf{P} means forming a vector containing the category mean for each case. Adjusting for the categorical variable means applying the residual mapping corresponding to \mathbf{P} , that is, $\mathbf{I} - \mathbf{P}$, which subtracts the category mean for each case.