

Problem Set 1 Solutions

Review of Linear Algebra and Probability

Prof. Jim Rehg
CS 7635 Computational Perception
College of Computing
Georgia Institute of Technology

January 14, 2002

Please make sure you understand all the solutions. In particular, problem 2 is relevant to our discussion of Principle Component Analysis and problem 4 is relevant to the tracking and human motion modeling material we will cover in the latter half of the semester.

Vectors and matrices are typeset in bold. A vector \mathbf{v} is a column vector; \mathbf{v}^T is a row vector where “T” denotes the transpose. $P(\cdot)$ denotes a probability density function (PDF), which sums or integrates to one.

1. [20 points]

(1a) The columns are linearly independent and the rank is 2 ($r = 2, m = 4, n = 2$). $r < m$ means there will not always be a solution for a given \mathbf{b} . $r = n$ means that any solution will be unique (the mapping \mathbf{A} is one-to-one but not onto).

(part ii) By inspection, \mathbf{b} lies in the column space of \mathbf{A} . The unique solution can be obtained by Gaussian Elimination as follows:

$$\begin{bmatrix} 1 & 3 & 10 \\ -4 & 2 & 12 \end{bmatrix} \Rightarrow \begin{bmatrix} 1 & 3 & 10 \\ 0 & 14 & 42 \end{bmatrix} \Rightarrow \mathbf{x} = \begin{bmatrix} -8/7 \\ 26/7 \end{bmatrix}.$$

More laboriously, the solution can be obtained from the pseudoinverse \mathbf{A}^+ :

$$\mathbf{A}^+ = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \text{ and } \mathbf{x} = \mathbf{A}^+ \mathbf{b}. \text{ Continuing,}$$
$$\mathbf{c} = \mathbf{A}^T \mathbf{b} = \begin{bmatrix} -38 \\ 54 \end{bmatrix}, \quad \mathbf{A}^T \mathbf{A} = \begin{bmatrix} 17 & -5 \\ -5 & 13 \end{bmatrix}, \quad \mathbf{x} = \frac{\begin{bmatrix} 13 & 5 \\ 5 & 17 \end{bmatrix}}{\det(\mathbf{A}^T \mathbf{A})} \begin{bmatrix} -38 \\ 54 \end{bmatrix}.$$

The last expression follows from the formula for the inverse of a 2x2 matrix.

(part iii) $\mathbf{y} \neq 0$ does not exist by uniqueness of \mathbf{x} . An orthonormal basis for the left null space is given by $\{[0010]^T, [0001]^T\}$. Any \mathbf{z} with a component in this space will destroy the existence of an exact solution. Any \mathbf{z} with a component in the column space will change the solution. So $\mathbf{z} \neq 0$ does not exist.

(1b) The columns are linearly dependent and the rank is 2 ($r = 2, m = 2, n = 4$). $r = m$ means that there will always be a solution for any \mathbf{b} . $r < n$ means that the solution will not be unique (the mapping \mathbf{A} is onto but not one-to-one).

(part ii) We can obtain the minimum norm solution from the pseudoinverse. But we can't use the construction procedure from (1a) because $r < n$ (and therefore $\mathbf{A}^T \mathbf{A}$ is singular). In general we can use the SVD to compute \mathbf{A}^+ , but in this case it's simpler to just use the null space.

A basis for the null space is constructed from $\mathbf{A}\mathbf{x} = \mathbf{0}$ by setting each of the free components¹ of \mathbf{x} to one, the remaining free components to zero, and solving

¹The free components of \mathbf{x} correspond to the columns of \mathbf{A} that do not contain pivots: x_2 and x_4 (the pivots are both 1's in this case)

for the remaining elements of \mathbf{x} :

$$\mathbf{A} \begin{bmatrix} x_1 \\ 1 \\ x_3 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Rightarrow \mathbf{u}_1 = \begin{bmatrix} -2 \\ 1 \\ 0 \\ 0 \end{bmatrix}; \quad \mathbf{A} \begin{bmatrix} x_1 \\ 0 \\ x_3 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Rightarrow \mathbf{u}_2 = \begin{bmatrix} -5 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

The minimum norm solution will have no component in the null space (i.e. $\mathbf{x}^T \mathbf{u}_1 = \mathbf{x}^T \mathbf{u}_2 = 0$). Since $x_3 = 1$ by inspection, we can set up a 3x3 linear system incorporating the first row of \mathbf{A} and the null space constraints and solve by elimination:

$$\begin{bmatrix} x_1 & x_2 & x_3 & b \\ 1 & 2 & 5 & -3 \\ -2 & 1 & 0 & 0 \\ -5 & 0 & 1 & 0 \end{bmatrix} \Rightarrow \begin{bmatrix} 1 & 2 & 5 & -3 \\ 0 & 5 & 10 & -6 \\ 0 & 0 & 6 & -3 \end{bmatrix} \Rightarrow \mathbf{x} = \begin{bmatrix} -1/10 \\ -1/5 \\ 1 \\ -1/2 \end{bmatrix}$$

(part iii) Since the dimension of the left null space is zero, $\mathbf{z} \neq 0$ does not exist. Any vector in the span of $\{\mathbf{u}_1, \mathbf{u}_2\}$ is a valid choice for \mathbf{y} .

(1c) The columns are linearly dependent and the rank is 1 ($r = 1, m = n = 2$). $r < m$ means there will not always be a solution for a given \mathbf{b} . $r < n$ means the solution, if it exists, will not be unique (the mapping \mathbf{A} is neither one-to-one nor onto).

The Singular Value Decomposition (SVD) of \mathbf{A} gives the expansion $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$. The linear system reduces to $\mathbf{\Sigma}\mathbf{w} = \mathbf{c}$, where $\mathbf{w} = \mathbf{V}^T\mathbf{x}$ and $\mathbf{c} = \mathbf{U}^T\mathbf{b}$. In this example, the SVD can be computed by inspection. A basis for the column space is $[3/5 \ 4/5]^T$. An orthonormal basis for the left null space is $[-4/5 \ 3/5]^T$. One singular value will be zero. Since the sum of the squares of the singular values equals the trace of $\mathbf{A}^T\mathbf{A}$, the other singular value is 25. We have

$$\mathbf{U} = \begin{bmatrix} 3/5 & -4/5 \\ 4/5 & 3/5 \end{bmatrix}, \quad \mathbf{\Sigma} = \begin{bmatrix} 25 & 0 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad \mathbf{c} = \begin{bmatrix} 25 \\ 5 \end{bmatrix}.$$

(part ii) Since \mathbf{b} has a nonzero projection into the left null space ($c_2 \neq 0$), the equation has no exact solution. The least squares solution is given by solving $25w_1 = 25$, yielding $w_1 = 1$. The minimum norm solution is obtained by setting $w_2 = 0$. A basis for the row space is given by $[4/5 \ 3/5]^T$. An orthonormal basis

for the null space is given by $[-3/5 \ 4/5]^T$. We have

$$\mathbf{V} = \begin{bmatrix} 4/5 & -3/5 \\ 3/5 & 4/5 \end{bmatrix}, \quad \mathbf{x} = \mathbf{V}\mathbf{w} = \begin{bmatrix} 4/5 \\ 3/5 \end{bmatrix}$$

(part iii) Any vector in the null space can be added to \mathbf{x} without changing its mapping into the column space (where the projection of \mathbf{b} lives). Thus we have

$$\mathbf{A}(\mathbf{x} + \alpha\mathbf{y}) = \mathbf{U} \begin{bmatrix} c_1 \\ 0 \end{bmatrix} = \mathbf{U} \begin{bmatrix} 25 \\ 0 \end{bmatrix} = \begin{bmatrix} 15 \\ 20 \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} -3/5 \\ 4/5 \end{bmatrix}.$$

Similarly, any vector in the left null space can be added to \mathbf{b} without changing the least squares solution \mathbf{x} . Let Σ^+ denote the pseudoinverse of Σ . Then

$$\mathbf{x} = \mathbf{V}\Sigma^+\mathbf{U}^T(\mathbf{b} + \alpha\mathbf{z}), \quad \Sigma^+ = \begin{bmatrix} 1/25 & 0 \\ 0 & 0 \end{bmatrix}, \quad \mathbf{z} = \begin{bmatrix} -4/5 \\ 3/5 \end{bmatrix}.$$

2. [20 points]

(2a)

$$P(\mathbf{x}) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \mu)^T \Sigma^{-1}(\mathbf{x} - \mu)\right\}$$

(2b) We want to “unwhiten” and “uncenter” the given data. We need to shift it by the mean and find \mathbf{B} such that $\mathbf{y} = \mathbf{B}\mathbf{x} + \mu$ has the desired covariance:

$$E\{(\mathbf{y} - \mu)(\mathbf{y} - \mu)^T\} = \Sigma \Rightarrow \mathbf{B}E\{\mathbf{x}\mathbf{x}^T\}\mathbf{B}^T = \mathbf{U}\mathbf{S}\mathbf{U}^T \Rightarrow \mathbf{B} = \mathbf{U}\mathbf{S}^{1/2},$$

where \mathbf{S} is a diagonal matrix containing the eigenvalues of Σ (which are positive since Σ is positive definite), and \mathbf{U} is a matrix with the associated eigenvectors as its columns (it is orthonormal since Σ is symmetric). The characteristic polynomial is $\lambda^2 - 3/4\lambda + 1/8$. It follows that:

$$\mathbf{S} = \begin{bmatrix} 1/4 & 0 \\ 0 & 1/2 \end{bmatrix}, \quad \mathbf{U} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}, \quad \text{and} \quad \mathbf{B} = \begin{bmatrix} 1/(2\sqrt{2}) & -1/2 \\ 1/(2\sqrt{2}) & 1/2 \end{bmatrix}$$

(2c) The equiprobability contours are centered at μ and indexed by the choice of k in: $\mathbf{y}^T \Sigma^{-1} \mathbf{y} = k$. Since Σ is positive definite, they are ellipses and their major and minor axes are given by the eigenvectors.

3. [20 points]

This is an example of inference in a graphical model (a Bayesian network in this case.) The network structure is that of a Naive Bayes classifier. The classifier node W has two conditionally independent measurements, C and M .

(3a) $P(C, W) = P(C|W)P(W) = P(M|W)P(W) = P(M, W)$. Note that in this equality what we are stating is that the two joint distributions are identical (same tables):

$$P(C, W) = \begin{bmatrix} 0.56 & 0.03 \\ 0.14 & 0.27 \end{bmatrix} = P(M, W)$$

Marginalizing over W in the above table yields

$$P(C) = P(M) = [0.59 \quad 0.41]$$

(3b) $C = 1$. Bayes Rule gives

$$P(W|C = 1) = \frac{P(C = 1|W)P(W)}{P(C = 1)} = \frac{1}{0.59} [0.8 \cdot 0.7 \quad 0.1 \cdot 0.3] = [0.95 \quad 0.05]$$

Note that the marginal distribution over C from part (b) supplies the correct normalizing factor above. Note also that the effect of the additional information is to increase the likelihood of $W = 1$ and decrease the likelihood of $W = 0$, as expected.

(3c) We have

$$P(M, W) = P(M|W)P(W) = \begin{bmatrix} 0.8 \cdot 0.95 & 0.1 \cdot 0.05 \\ 0.2 \cdot 0.95 & 0.9 \cdot 0.05 \end{bmatrix} = \begin{bmatrix} 0.76 & 0.005 \\ 0.19 & 0.045 \end{bmatrix}$$

Marginalizing W out of $P(W, M)$ yields

$$P(M) = [0.765 \quad 0.235]$$

(3d) $W = 0$. By the definition of conditional independence,

$$P(M|W = 0) = [0.1 \quad 0.9]$$

4. [20 points]

(4a) Note that there is a typo in the problem set: The two variances are σ_x^2 and σ_n^2 . Also, it is implicit that x and n are independent, but this should have been stated explicitly. Since y is the sum of two Gaussian random variables it is also Gaussian. We have $y \sim N(\mu, \sigma_x^2 + \sigma_n^2)$ and $p(y|x) = N(x, \sigma_n^2)$. Applying Bayes rule:

$$\begin{aligned} p(x|y) &= \frac{p(y|x)p(x)}{p(y)} = \frac{1}{p(y)2\pi\sigma_x^2\sigma_n^2} \exp \left\{ -\frac{1}{2} \left[\frac{(y-x)^2}{\sigma_n^2} + \frac{(x-\mu)^2}{\sigma_x^2} \right] \right\} \\ &= K \exp \left\{ -\frac{1}{2} \left(x - \frac{\sigma_x^2 y + \sigma_n^2 \mu}{\sigma_x^2 + \sigma_n^2} \right)^2 \left(\frac{\sigma_x^2 + \sigma_n^2}{\sigma_x^2 \sigma_n^2} \right)^{-1} \right\}. \end{aligned}$$

The last equality follows from completing the square in x and lumping the constant terms left in the exponential into K along with $p(y)$. Since x is the only random variable in the expression, y and μ are constants. For least squares estimation, K is just a normalizing constant and does not need to be evaluated. It can be found in standard texts² for the case where $\mu = 0$:

$$K = \left(2\pi \frac{\sigma_n^2 \sigma_x^2}{\sigma_n^2 + \sigma_x^2} \right)^{-1}$$

Continuing, we want to find the \hat{X} that minimizes:

$$\begin{aligned} E\{(x - \hat{X})^2|y\} &= \int_{-\infty}^{+\infty} (x - \hat{X})^2 p(x|y) dx \\ &= \left(\hat{X} - \int_{-\infty}^{+\infty} x p(x|y) dx \right)^2 + \int_{-\infty}^{+\infty} x^2 p(x|y) dx - \left(\int_{-\infty}^{+\infty} x p(x|y) dx \right)^2. \end{aligned}$$

The last equation follows from completing the square in \hat{X} . The minimum is attained by the conditional mean, and the remaining nonzero terms give the expected mean square error. Summarizing:

$$\hat{X} = E\{x|y\} = \frac{\sigma_x^2 y + \sigma_n^2 \mu}{\sigma_x^2 + \sigma_n^2} \quad \text{and} \quad E\{(x - \hat{X})^2|y\} = E\{x^2|y\} - \hat{X}^2$$

²See for example *Optimal Filtering* by B. D. O. Anderson and J. B. Moore, Prentice Hall, 1979, p. 24.

(4b) Note: It is implicit in the problem statement that the y_i 's are zero mean, but this should have been stated explicitly. We want to find the \mathbf{a} such that $\hat{s} = \mathbf{y}^T \mathbf{a}$ is the minimum mean square estimate of s . Thus \mathbf{a} should be chosen to minimize the mean square error $e = E\{(s - \hat{s})^2\} = E\{(s - \mathbf{y}^T \mathbf{a})^2\}$. Differentiating e with respect to \mathbf{a} yields a system of linear equations which can be solved by elimination:

$$E\{\mathbf{y}(s - \mathbf{y}^T \mathbf{a})\} = \mathbf{0} \quad \Rightarrow \quad E\{\mathbf{y}\mathbf{y}^T\} \mathbf{a} = E\{s\mathbf{y}\} \quad \Rightarrow \quad \mathbf{a} = \begin{bmatrix} 2 \\ 1 \\ 2 \end{bmatrix}.$$

The linear system above is known as the *Yule-Walker equations*. It was introduced by Yule in his seminal study of sunspot data: G. U. Yule. "On a method of investigating periodicities in disturbed series with special reference to Wolfer's sunspot numbers." *Philosophical Trans. of Royal Society of London A*, 226, pages 267-298, 1927.

The Yule-Walker equations have an important geometric interpretation. To develop it, we first have to establish that a collection of n zero mean random variables (in this case the y_i 's making up the vector \mathbf{y}) can be viewed as "vectors" in a certain abstract inner product space. Let \mathcal{V} denote the linear subspace spanned by the elements of \mathbf{y} . Each element of \mathcal{V} (i.e. each "vector") has the form $\sum_i \alpha_i y_i$ for some choice of real α_i 's. It is easy to show that \mathcal{V} forms a vector space of dimension n .³ We define $E\{uv\}$ to be the inner product for two vectors $u, v \in \mathcal{V}$.⁴ Then the squared "norm" of a random variable u is its variance $E\{u^2\}$. When $E\{uv\} = 0$ we say that u and v are orthogonal.

We are now ready to interpret the Yule-Walker equations as a statement of the following *orthogonality principle*: The minimum mean square estimator of a random variable s has the property that the error $\varepsilon = s - \hat{s}$ is *orthogonal* to the data y_i . Each of the n Yule-Walker equations is a statement of orthogonality:

$$E\{(s - \hat{s})y_i\} = E\{\varepsilon y_i\} = 0$$

We can take arbitrary linear combinations of these equations to obtain $E\{\varepsilon(\sum_i \alpha_i y_i)\} = 0$. It follows that ε is orthogonal to \mathcal{V} . We can interpret \mathcal{V} as the space of all possi-

³Let u, v be two elements of \mathcal{V} with coordinates $\{\alpha_i\}, \{\beta_i\}$ respectively. Then for arbitrary a, b we have $au + bv = \sum_i (a\alpha_i + b\beta_i)y_i \in \mathcal{V}$. Furthermore, $\{y_i\}$ is an n -dimensional basis for \mathcal{V} since it spans the space by construction and is linearly independent (because $\sum_i \lambda_i y_i = 0 \Rightarrow \lambda_i = 0 \forall i$ by uniqueness of 0.)

⁴This is well-defined, since $E\{(au + bv)w\} = aE\{uw\} + bE\{vw\}$ and $E\{u^2\} \geq 0$, with equality only when $u = 0$.

ble linear estimators of s . The optimal linear estimator is given by the *projection* of s onto \mathcal{V} . As a consequence, this result is also known as the *projection theorem*.

The projection theorem is the cornerstone of the ARMA (autoregressive moving average) approach to time series modeling. For example, the y_i 's could be the last n values in a time series and s could be the next value in the series which we would like to predict. The projection theorem is also the key insight behind the Kalman filter, which addresses optimal estimation in state space models. It leads directly to the concept of the innovation and orthogonal increments.

5. [Extra Credit: +5 points]

We need to prove

$$(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) = \text{tr}\{\boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T\}.$$

The result can be established by means of the following identity for n -vectors \mathbf{y} and \mathbf{z} ,

$$\mathbf{y}^T \mathbf{z} = \text{tr}\{\mathbf{y} \mathbf{z}^T\} = \text{tr}\{\mathbf{z} \mathbf{y}^T\},$$

by setting $\mathbf{y} = \mathbf{x} - \boldsymbol{\mu}$ and $\mathbf{z} = \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$.