CSci 6974 and ECSE 6966 Math. Tech. for Vision, Graphics and Robotics
Lecture 8, February 13, 2006
SVD and Other Decompositions

Overview of Today's Lecture

As one application of linear algebra and to provide background on estimation, Lecture 8 focuses on estimating a line (plane) from a set of points. We'll consider both *ordinary regression* and *orthogonal regression*.

Ordinary Regression — Definition

- The measured points are of the form $(\mathbf{x}_i^\top, y_i)^\top$, where $\mathbf{x}_i$ is the location at which a measurement is made and y_i is the measurement itself.
  - In general we write $\mathbf{x}^\top = (x_1, \ldots, x_n)$.
  - In two-dimensions, the vector $\mathbf{x}_i$ just becomes the scalar x_i.
  - If you think of an intensity image as a function $z = I(x, y)$, then the pixel coordinate vectors $(x, y)^\top$ play the role of $\mathbf{x}_i$ in the point measurement and the intensity value z plays the role of y_i.
- The line (plane or hyperplane) to be estimated is described by the equation

$$y = a_0 + a_1 x_1 + a_2 x_2 + \cdots + a_n x_n \tag{1}$$

  In vector form this can be written in a number of equivalent ways:

$$y = \tilde{\mathbf{a}}^\top \mathbf{x} + a_0 \tag{2}$$

  where $\tilde{\mathbf{a}}^\top = (a_1, \ldots, a_n)$, or

$$y = \left(\mathbf{x}^\top, 1\right) \mathbf{a} \tag{3}$$

  where $\mathbf{a}^\top = (\tilde{\mathbf{a}}^\top, a_0) = (a_1, \ldots, a_n, a_0)$., or

$$y = \mathbf{a}^\top \begin{pmatrix} \mathbf{x} \\ 1 \end{pmatrix}. \tag{4}$$

  - Note that if $\mathbf{x}_i$ has n dimensions, then $\mathbf{a}$ has $n + 1$.
  - When we write $\mathbf{a}^\top = (\tilde{\mathbf{a}}^\top, a_0)$, we are distinguishing a_0 as a special component — the "intercept" component — of $\mathbf{a}$.
- For any data point $(\mathbf{x}_i^\top, y_i)$ and for any possible parameter vector $\mathbf{a}$, the error is

$$e_i = y_i - \left(\mathbf{x}_i^\top, 1\right)\mathbf{a}. \tag{5}$$

  This is the difference between the measured value at $\mathbf{x}_i$ and the plane equation's value at $\mathbf{x}_i$.

- Our goal for a given set of data points is to find the parameter vector $\mathbf{a}$ that minimizes a function of the e_i values.

Choice of Error Functions

- Historically, two choices of error functions were first considered:
  - L_1 or absolute norm:

$$E_1(\mathbf{a}) = \sum_i |e_i| \;=\; \sum_i |y_i - \big(\mathbf{x}_i^\top, 1\big)\mathbf{a}| \tag{6}$$

  - L_2 or least-squares norm:

$$E_2(\mathbf{a}) = \sum_i [e_i^2] = \sum_i \big[y_i - \big(\mathbf{x}_i^\top, 1\big)\mathbf{a}\big]^2 \tag{7}$$

- The L_2 norm gained favor because it is computational very simple and produces the optimal estimate for "normally" (Gaussian) distributed errors.
- When errors don't follow a normal distribution, the L_1 norm is better. BUT, other norms, referred to as "robust norms" are even better, so the L_1 norm is not often used. We will study robust norms much later in the semester.

Ordinary Regression Estimate — Component Form

As the simplest possible starting point, we consider the case of fitting the regression line $y = a_1 x + a_0$ in two dimensions. We start using component notation.

- Keep in mind throughout the discussion that the N points (x_i, y_i) are known and the line parameters are not known.
- The objective function is

$$E(a_0, a_1) = \sum_i (y_i - a_1 x_i - a_0)^2. \tag{8}$$

- Our goal is to minimize this quadratic function of a_0 and a_1, and thereby find the "estimated" a_0 and a_1.
- Calculating the derivative of E with respect to a_0 and a_1, setting each derivative equal to 0, writing the result in matrix form, and solving yields

$$\hat{\mathbf{a}} = \begin{pmatrix} \hat{a}_1 \\ \hat{a}_0 \end{pmatrix} = \begin{pmatrix} \sum_i x_i^2 & \sum_i x_i \\ \sum_i x_i & N \end{pmatrix}^{-1} \begin{pmatrix} \sum_i x_i y_i \\ \sum_i y_i \end{pmatrix}. \tag{9}$$

Regression Estimate — Vector Form

- In vector form the linear regression objective function becomes

$$E(\mathbf{a}) = \sum_i \left[y_i - \left(\mathbf{x}_i^\top, 1 \right) \mathbf{a} \right]^2. \tag{10}$$

  There is no explicit dependence on dimension in this equation, so the following derivation applies equally to lines, plane, and hyperplanes.

- The minimization is accomplished by solving

$$\frac{\partial E}{\partial \mathbf{a}} = \mathbf{0}. \tag{11}$$

- Why do we know that the answer will be a minimum?

Aside: Vector and Matrix Derivatives

Pointers to resources for the computation of vector and matrix derivatives will be provided on the course web page as well as in a separate class handout.

- The derivative of a scalar product, $\mathbf{a}^\top \mathbf{x} = \mathbf{x}^\top \mathbf{a}$, with respect to $\mathbf{a}$ is just $\mathbf{x}$.

- More generally, the derivative of a function of a scalar product, $f(\mathbf{a}^\top \mathbf{x})$ is

$$f'(\mathbf{a}^\top \mathbf{x})\mathbf{x}. \tag{12}$$

- The derivative of a product $\mathbf{X}\mathbf{a}$ with respect to the vector $\mathbf{a}$ is just $\mathbf{X}$.

- The derivative of a quadratic form $\mathbf{a}^\top \mathbf{X} \mathbf{a}$ is

$$\mathbf{X}\mathbf{a} + \mathbf{X}^\top \mathbf{a}. \tag{13}$$

  Fortunately, in most cases we're interested in, $\mathbf{X}$ will be symmetric, so this reduces to

$$2\mathbf{X}\mathbf{a}. \tag{14}$$

- These rules, in combination with the chain rule, are pretty-much all you will need.

Back to the Vector Form

- Taking the derivative of

$$E(\mathbf{a}) = \sum_i \left[y_i - \left(\mathbf{x}_i, 1 \right)^\top \mathbf{a} \right]^2 \tag{15}$$

  with respect to $\mathbf{a}$, setting the result to $\mathbf{0}$, and solving yields

$$\hat{\mathbf{a}} = \left(\sum_i \begin{pmatrix} \mathbf{x}_i \\ 1 \end{pmatrix} \begin{pmatrix} \mathbf{x}_i^\top & 1 \end{pmatrix} \right)^{-1} \begin{pmatrix} \sum_i y_i \mathbf{x}_i \\ \sum_i y_i \end{pmatrix} \tag{16}$$

Regression Estimate — Matrix Form

Writing

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^\top & 1 \\ \mathbf{x}_2^\top & 1 \\ \mathbf{x}_3^\top & 1 \\ \cdots & \cdots \\ \mathbf{x}_N^\top & 1 \end{pmatrix} \qquad \text{and} \qquad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \cdots \\ y_n \end{pmatrix}, \tag{17}$$

the problem becomes solving the over-constrained system

$$\mathbf{X}\mathbf{a} = \mathbf{y}. \tag{18}$$

- The least-squares objective becomes

$$E(\mathbf{a}) = \big(\mathbf{X}\mathbf{a} - \mathbf{y}\big)^\top \big(\mathbf{X}\mathbf{a} - \mathbf{y}\big). \tag{19}$$

  The minimization is accomplished by solving

$$\frac{\partial E}{\partial \mathbf{a}} = \mathbf{0}. \tag{20}$$

- Working through the math, this yields

$$\hat{\mathbf{a}} = \big(\mathbf{X}^\top \mathbf{X}\big)^{-1} \mathbf{X}^\top \mathbf{y}, \tag{21}$$

  assuming the inverse exists.

Regression Estimate — SVD

We can solve the minimization problem in the matrix form directly from the SVD of $\mathbf{X}$.

- Let the SVD of $\mathbf{X}$ be $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$.
- Then $\mathbf{X}^\top\mathbf{X} = \mathbf{V}\mathbf{D}^2\mathbf{V}^\top$.
- Hence, $(\mathbf{X}^\top\mathbf{X})^{-1} = \mathbf{V}\mathbf{D}^{-2}\mathbf{V}^\top$.
- Finally,

$$\hat{\mathbf{a}} = \big(\mathbf{X}^\top \mathbf{X}\big)^{-1} \mathbf{X}^\top \mathbf{y} = \mathbf{V}\mathbf{D}^{-1}\mathbf{U}^\top \mathbf{y}. \tag{22}$$

- Notice that we never had to compute $\mathbf{X}^\top\mathbf{X}$ explicitly.

Rank Deficiency

What if there are not enough constraints?

- This can appear when there are too few points, or when there are enough points, but they lie in a proper subspace of $\mathbb{R}^n$.

- This shows up as $\mathbf{X}$ being of rank less than $n+1$.
- In the SVD of $\mathbf{X}^\top\mathbf{X}$ (or just of $\mathbf{X}$), one or more singular values will be 0 or near 0.
- Among the infinite solutions, when one must be chosen the one preferred is the one that sits entirely in the row-space (no component in the nullspace).
- If there are k non-zero singular values, $\sigma_1, \ldots, \sigma_k$ (remember that by convention these are ordered from greatest to smallest in the SVD), we write the matrix
$$\mathbf{D}^+ = \text{diag}(1/\sigma_1, \ldots, 1/\sigma_k, 0, \ldots, 0) \tag{23}$$
and compute
$$\hat{\mathbf{a}} = \mathbf{V}\mathbf{D}^+\mathbf{U}^\top\mathbf{y} = \sum_{i=1}^{k} \frac{1}{\sigma_i}\mathbf{v}_i\mathbf{u}^\top\mathbf{y} \tag{24}$$
- This is also the "shortest" solution.
- More generally,
$$\mathbf{X}^+ = \mathbf{V}\mathbf{D}^+\mathbf{U}^\top \tag{25}$$
is known as the *pseudo-inverse* of $\mathbf{X}$.
- In practice, the choice of whether to use the pseudo-inverse solution or to report that no solution exists will depend on the context.

Orthogonal Regression Problem

- Given N points $\mathbf{x}_i$ in $\mathbb{R}^n$, we want to find the (hyper)plane closest to the points.
- This plane will have $n+1$ parameters, satisfying the implicit equation:
$$\big(\mathbf{x}^\top, 1\big)\mathbf{a} = 0. \tag{26}$$
- Notice that scaling $\mathbf{a}$ by a constant does not change this equation, so there are infinitely many parameter vectors. We'll address this soon.
- Since we no longer distinguish a special coordinate as the measurement direction all coordinates must be treated equally in the error measure, e_i.

The Error Distance

- The error distance, e_i, between a data point $\mathbf{x}_i$ and a plane described by parameter vector $\mathbf{a}_i$ is defined as the distance between $\mathbf{x}_i$ and the closest point to $\mathbf{x}_i$ satisfying the plane constraint equation.
- Fortunately, we know that the line joining $\mathbf{x}_i$ and the closest point must be normal to the plane. We can use this to derive a measure of the distance.

- Writing $\mathbf{a}^\top = (\tilde{\mathbf{a}}^\top, a_0)^\top$, as above, but meaning something slightly different, we can show that the distance from $\mathbf{x}_i$ to the plane is

$$\frac{\big(\mathbf{x}_i^\top, 1\big)\mathbf{a}}{\tilde{\mathbf{a}}^\top} \tag{27}$$

- The vector $\tilde{\mathbf{a}}$ is normal to the plane. Ensuring that this is a unit vector — something we must enforce — we have the distance

$$e_i = \big(\mathbf{x}_i^\top, 1\big)\mathbf{a} \tag{28}$$

  We'll call this the normal or orthogonal error distance

Minimizing the Orthogonal Error

- Gather all data points into a matrix:

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^\top & 1 \\ \mathbf{x}_2^\top & 1 \\ \cdots & \\ \mathbf{x}_N^\top & 1 \end{pmatrix} \tag{29}$$

- Our goal is to find the $\mathbf{a}$ that minimizes

$$\|\mathbf{X}\mathbf{a}.\| \tag{30}$$

- The first big problem we face is that the trivial solution $\mathbf{a} = \mathbf{0}$ minimizes this.

- Since $\mathbf{a}$ is defined only up to a scale factor and since the trivial solution $\mathbf{a} = \mathbf{0}$ is not allowed, we need to impose a constraint on $\mathbf{a}$. The simplest one is $\mathbf{a}^\top\mathbf{a} = 1$.

  – This doesn't quite enfore the normal to be a unit vector. More on this soon.

- Our problem then become to minimize

$$\|\mathbf{X}\mathbf{a}\| \qquad \text{subject to} \qquad \mathbf{a}^\top\mathbf{a} - 1 = 0. \tag{31}$$

- We will show in class that this solution is the last column of $\mathbf{V}$ in the SVD of $\mathbf{X}$.

- Conversion to a unit normal is then straightforward.

Practice Problems / Potential Test Questions

1. We gave three different derivations for the solution to the ordinary regression problem. Prove that they produce the same answers. Do this by converting all of the equations to component form.

2. The regression error measure is

$$e_i = y_i - \big(\mathbf{x}_i^\top, 1\big)\mathbf{a},$$

and the orthogonal regression error measure is

$$e_i' = \big(\mathbf{z}_i^\top, 1\big)\mathbf{b}$$

Derive an expression relating these two errors. In other words, translate the y_i and $\mathbf{x}_i$ vector into $\mathbf{z}_i$, expression $\mathbf{a}$ in terms of $\mathbf{b}$, and then derive a relationship between e_i' and e_i.

Problems For Grading

Submit solutions to the following problems on Tuesday, February 21st, as part of HW 5. Graded HW papers will be available outside my office door by Wednesday morning, February 22nd.

1. (**15 points**) Show in detail how to find the single point in $\mathbb{R}^2$ closest to n lines with parameter vectors (a_i, b_i, c_i). Remember, in $\mathbb{R}^2$, each line is

$$a_i x + b_i y + c_i = 0,$$

and you may assume that $a_i^2 + b_i^2 = 1$. Your goal is to develop a least-squares solution that produces the values of x and y of the point.