EE 159/CS 168 - Convex Optimization
Scott Fullenbaum

Homework 2

---

1. First, we can define our second regression equation as $\widetilde{y} = \widetilde{x}^T\beta + v$. We can take the difference of this with our original equation getting:
   $|\hat{y} - \widetilde{y}| = |x^T\beta + v - (\widetilde{x}^T\beta + v)| = |x^T\beta - \widetilde{x}^T\beta|$. Since this is a scalar, we can take the transpose of the right side, getting $|\beta^T x^T - \beta^T\, \widetilde{x}| = |\beta^T(x^T - \widetilde{x})|$. This is a dot product of two vectors, so we can use the Cauchy-Schwarz inequality and $|\hat{y} - \widetilde{y}| = |\beta^T(x^T - \widetilde{x})| \leq \|\beta\|\|x - \widetilde{x}\|$, which proves: $|\hat{y} - \widetilde{y}| \leq \|\beta\|\|x - \widetilde{x}\|$

2. $\frac{\partial Q}{\partial x} = 2ax + 4y + 2az$, $\frac{\partial Q}{\partial y} = 8ay + 4x + 4z$, $\frac{\partial Q}{\partial z} = 8az + 2ax + 4y$.
   $\nabla Q = \begin{bmatrix} 2ax + 4y + 2az \\ 8ay + 4x + 4z \\ 8az + 2ax + 4y \end{bmatrix}$. To find the Hessian, we need to find the second partials, and can use the fact that the order in which they are taken doesn't matter:
   $\frac{\partial Q}{\partial x \partial x} = 2a$ $\frac{\partial Q}{\partial x \partial y} = 4$ $\frac{\partial Q}{\partial x \partial z} = 2a$ $\frac{\partial Q}{\partial y \partial y} = 8a$ $\frac{\partial Q}{\partial y \partial z} = 4$. $\nabla^2 Q = \begin{bmatrix} 2a & 4 & 2a \\ 4 & 8a & 4 \\ 2a & 4 & 8a \end{bmatrix}$

   To see if the Hessian is positive semi definite, we can use Sylvester's criterion and evaluate the determinant of the upper $1x1$ matrix, $2x2$ matrix, and determinant of the matrix is ≥ 0.
   $1x1$: det(2a) $= 2a \geq 0$ for $a \geq 0$
   $2x2$: $\begin{vmatrix} 2a & 4 \\ 4 & 8a \end{vmatrix} = 16a^2 - 16 \geq 0$ for $a \leq -1$ or $a \geq 1$
   $3x3$: 2a$\begin{vmatrix} 8a & 4 \\ 4 & 8a \end{vmatrix}$ - 4$\begin{vmatrix} 4 & 4 \\ 2a & 8a \end{vmatrix}$ + 2a$\begin{vmatrix} 4 & 8a \\ 2a & 4 \end{vmatrix}$ $= 96a^3 - 96a \geq 0$ for $-1 \leq a \leq 0$ or $a \geq 1$
   Combining all three restrictions, for all to be positive, $a \geq 1$, meaning the Hessian is PSD if $a \geq 1$.

3. (a) $J(z) = \sum_{i=1}^{L} \|x_i - z\|^2 = \sum_{i=1}^{L} \|x_i - \bar{x} + \bar{x} - z\|^2 = \sum_{i=1}^{L} \|(x_i - \bar{x}) - (z - \bar{x})\|^2 =$
   $\sum_{i=1}^{L} \langle (x_i - \bar{x}) - (z - \bar{x}), (x_i - \bar{x}) - (z - \bar{x}) \rangle = \sum_{i=1}^{L} \langle x_i - \bar{x}, x_i - \bar{x} \rangle - 2\langle x_i - x, z - \bar{x} \rangle + \langle z - \bar{x}, z - \bar{x} \rangle$
   $= \sum_{i=1}^{L} \|x_i - \bar{x}\|^2 - 2(x_i - x)^T(z - \bar{x}) + \|z - \bar{x}\|^2 = \sum_{i=1}^{L} (\|x_i - \bar{x}\|^2 - 2(x_i - x)^T(z - \bar{x})) + L\|z - \bar{x}\|^2$.
   The last step can be made as $\|z - \bar{x}\|^2$ is constant.

   (b) $\sum_{i=1}^{L} (x_i - \bar{x})^T(z - \bar{x}) = \sum_{i=1}^{L} ((x_i - \bar{x})^T) * (z - \bar{x}) = (\sum_{i=1}^{L} (x_i - \bar{x}))^T * (z - \bar{x})$.
   $\sum_{i=1}^{L} x_i - \bar{x} = \sum_{i=1}^{L} x_i - L\bar{x} = \sum_{i=1}^{L} x_i - L\frac{1}{L}\sum_{i=1}^{L} x_i = \sum_{i=1}^{L} x_i - \sum_{i=1}^{L} x_i = 0$, so this equation is equal to 0.

(c) As $J(z) = \sum_{i=1}^{L}(\|x_i - \bar{x}\|^2 - 2(x_i - x)^T(z - \bar{x})) + L\|z - \bar{x}\|^2 = \sum_{i=1}^{L}\|x_i - \bar{x}\|^2 + L\|z - \bar{x}\|^2$. $L\|z - \bar{x}\|^2 \geq 0$, and is equal to 0 only when $z = \bar{x}$, meaning $L\|z - \bar{x}\|^2 > 0$ for $z \neq \bar{x}$, and as the term in the summation is constant for any z, then $J(z) > J(\bar{x})$ for $z \neq \bar{x}$ and $z = \bar{x}$ minimizes J(z).

4. To find the gradient and Hessian, and can use the chain rule, rewriting $f(x)$ as $f(x) = g(h(x))$ where $h(x) = 1 + \|Ax + b\|_2^2$ and $g(y) = log(y)$.

Gradient: $\nabla f(x) = g'h(x)\nabla h(x) = \frac{1}{1 + \|Ax + b\|_2^2}\nabla h(x) = \frac{2A^TAx + 2A^Tb}{1 + \|Ax + b\|_2^2}$

Hessian: $\nabla^2 f(x) = g''(h)\nabla h(x)\nabla^T h(x) + g'(h)\nabla^2 h(x) =$

$\frac{-1}{(1 + \|Ax + b\|_2^2)^2} * (2A^TAx + 2A^Tb)(2A^TAx + 2A^Tb)^T + \frac{2A^TA}{1 + \|Ax + b\|_2^2}$

$= \frac{-(2A^TAx + 2A^Tb)(2A^TAx + 2A^Tb)^T}{(1 + \|Ax + b\|_2^2)^2} + \frac{2A^TA}{1 + \|Ax + b\|_2^2}$

5. (a) Start with the likelihood function:

$l(R, a) = -(Nn)/2log(2\pi) - (N/2)logdet(R) - 1/2\sum_{k=1}^{N}(y_k - a)^T R^{-1}(y_k - a)$.

Since most of the simplification is done with the third term, I will be ignoring the first two until the end. Since the sum is constant, the trace of it is equal to itself, so we can rewrite as:

$-1/2\sum_{k=1}^{N} tr((y_k - a)^T R^{-1}(y_k - a)) = -1/2\sum_{k=1}^{N} tr((y_k - a)(y_k - a)^T R^{-1})$

$= -1/2tr(\sum_{k=1}^{N}(y_k - a)(y_k - a)^T R^{-1})$. Now, let $y_k - a = ((y_k - \mu) - (a - \mu))$, so our sum is:

$= -1/2tr(\sum_{k=1}^{N}((y_k - \mu) - (a - \mu))((y_k - \mu) - (a - \mu))^T R^{-1})$

$= -1/2tr((\sum_{k=1}^{N}(y - y_k)(y - y_k)^T + N(a - \mu)(a - \mu)^T)R^{-1})$

There is a step between these two, involving terms in a matrix by $-2(y_k - \mu)(a - \mu)^T$. However, it gets a bit messy to type out and using the same logic from question 3.b, which holds for each component, all the terms in that matrix are 0. Also, as $(a - \mu)(a - \mu)^T$ is constant across the summation, we can just multiply itself by N. Now note $\sum_{i=1}^{N}(y - y_k)(y - y_k)^T = NY$, so the equation is:

$= -1/2(tr(NY + N(a - \mu)(a - \mu)^T)R^{-1})$

Distributing and using the property of traces that $tr(A + B) = tr(A) + tr(B)$ we can rewrite this sum as:

$= -1/2(tr(NYR^{-1}) + tr(N(a - \mu)(a - \mu)^T R^{-1})$.

Since $tr(cB)$ where c is a scalar $= c{*}tr(B)$ and $tr(AB) = tr(BA)$, the equation becomes:

$= -1/2(N * tr(R^{-1}Y) + N * tr((a - \mu)(a - \mu)^T R^{-1}))$

$= -1/2(N * tr(R^{-1}Y) + N * tr((a - \mu)^T R^{-1}(a - \mu)))$

Since the second term is a scalar, and $tr(c) = c \forall c \in R$

$= \frac{N}{2}(tr(R^{-1}Y) - (a-\mu)^T R^{-1}(a-\mu))$

Combining this with the other two terms and factoring out $N/2$ from them gives:

$l(R,a) = \frac{N}{2}(-nlog(2\pi) - logdet(R) - tr(R^{-1}Y) - (a-\mu)^T R^{-1}(a-\mu))$

(b) First, $\nabla_a = \frac{-N}{2}2R^{-1}(a-\mu) = -NR^{-1}(a-\mu)$ and $\nabla_a^2 = -NR^{-1}$

To find ∇_R, ignoring terms from the likelihood without R, the equation is

$l(R,a) = \frac{N}{2}(-logdetR - tr(R^{-1}Y) - (a-\mu)^T R^{-1}(a-\mu))$, using the table we get:

$\nabla_R = \frac{N}{2}((-R^{-1})^T - (-R^{-1}YR^{-1})^T)$

To find the ML estimates, set ∇_a and $\nabla_R = 0$.

For ∇_a, $-NR^{-1}(a-\mu) = 0$, so a unique minimum is reached at $a = \mu$.

For R, $\nabla_R = \frac{N}{2}((-R^{-1})^T - (-R^{-1}YR^{-1})^T) = 0$. Since $Y \succ 0$, Y is invertible, so there exists a unique matrix, Y^{-1} such that $YY^{-1} = I$ that is it's inverse. If we let $Y = R$, $\nabla_R = (-Y^{-1})^T - (-Y^{-1}YY^{-1})^T) = (-Y^{-1})^T - (-Y^{-1})^T = \mathbf{0}$

So, the ML estimates are $a = \mu$ and $R = Y$. They are unique as $(a-\mu)$ has one minimum, and there is only one inverse Y as it is invertible.