Directional derivatives for matrix calculus

Benjamin Chu

Graduate program in Biomathematics
Department of Computational Medicine
University of California, Los Angeles

January 28, 2021

Introduction

- In statistics, we often want to maximize things, requiring gradients and Hessians
- One way to obtain gradients and Hessians is through directional derivatives
- Compared to traditional matrix calculus, its primary advantages are:
  1. Gradient and Hessian are implicit in the 1st and 2nd directional derivatives
  2. Variables remain intact until gradients/Hessians are explicitly needed (e.g. $\boldsymbol{X}, \boldsymbol{Y}$ stay as $\boldsymbol{X}, \boldsymbol{Y}$ instead of $\mathrm{vec}(\boldsymbol{X})$...etc)
  3. Scalar/vector/matrix valued functions have scalar/vector/matrix valued directional derivatives

Definition

The (Hadamard semi-) directional derivative of $f(\boldsymbol{x})$ in the direction $\boldsymbol{v}$ is the limit

$$d_{\boldsymbol{v}} f(\boldsymbol{x}) = \lim_{\substack{h \to 0 \\ \boldsymbol{w} \to \boldsymbol{v}}} \frac{f(\boldsymbol{x} + h\boldsymbol{w}) - f(\boldsymbol{x})}{h}.$$

Directional derivatives enjoys the chain, sum, and product rules (Proposition 3.2.4 of MM optimization)

$$\begin{aligned} d_{\boldsymbol{v}}[f \circ g(\boldsymbol{x})] &= d_{d_{\boldsymbol{v}} g(\boldsymbol{x})} f[g(\boldsymbol{x})] \\ d_{\boldsymbol{v}}[f(\boldsymbol{x}) + g(\boldsymbol{x})] &= d_{\boldsymbol{v}} f(\boldsymbol{x}) + d_{\boldsymbol{v}} g(\boldsymbol{x}) \\ d_{\boldsymbol{v}}[f(\boldsymbol{x}) g(\boldsymbol{x})] &= d_{\boldsymbol{v}} f(\boldsymbol{x}) g(\boldsymbol{x}) + f(\boldsymbol{x}) d_{\boldsymbol{v}} g(\boldsymbol{x}) \end{aligned}$$

Furthermore, for a differentiable function $f(\boldsymbol{x})$, we have $d_{\boldsymbol{v}} f(\boldsymbol{x}) = df(\boldsymbol{x})\boldsymbol{v}$, where $df(\boldsymbol{x}) = \nabla f(\boldsymbol{x})^t$ is the first differential of $f(\boldsymbol{x})$. For a differentiable function $f(\boldsymbol{x}, \boldsymbol{y})$ of two variables, we also have $d_{(\boldsymbol{u},\boldsymbol{v})} f(\boldsymbol{x}, \boldsymbol{y}) = d_{(\boldsymbol{u},0)} f(\boldsymbol{x}, \boldsymbol{y}) + d_{(0,\boldsymbol{v})} f(\boldsymbol{x}, \boldsymbol{y})$.

Second Differentials

The second directional derivative is similarly defined

$$d_{\boldsymbol{u}}[d_{\boldsymbol{v}} f(\boldsymbol{x})] = \lim_{\substack{h \to 0 \\ \tilde{\boldsymbol{w}} \to \boldsymbol{u}}} \frac{d_{\boldsymbol{v}} f(\boldsymbol{x} + h\tilde{\boldsymbol{w}}) - d_{\boldsymbol{v}} f(\boldsymbol{x})}{h}.$$

For $f(\boldsymbol{x})$ twice differentiable, the second differential $d_{\boldsymbol{u}}[d_{\boldsymbol{v}} f(\boldsymbol{x})]$ is linear in $\boldsymbol{u}$ for $\boldsymbol{v}$ fixed and linear in $\boldsymbol{v}$ for $\boldsymbol{u}$ fixed. It is accordingly a bilinear form. When we set $\boldsymbol{u} = \boldsymbol{v}$, we get a quadratic form $d^2_{\boldsymbol{v}} f(\boldsymbol{x})$. At this juncture we have not yet associated a matrix to the quadratic form. In fact, if $\boldsymbol{x}$ is a matrix, then the matrix is replaced by a tensor. We can identify a matrix if we vectorize $\boldsymbol{x}$ by stacking its columns into a matrix.

Extracting gradients from directional derivatives

For vector $\boldsymbol{x}$, we can extract gradients and Hessians for twice differentiable functions as

$$d_{\boldsymbol{v}} f(\boldsymbol{x}) = \nabla f(\boldsymbol{x}) \cdot \boldsymbol{v} = (\nabla f(\boldsymbol{x}))^t \boldsymbol{v}$$
$$d^2_{\boldsymbol{v}} f(\boldsymbol{x}) = \boldsymbol{v}^t \boldsymbol{H} \boldsymbol{v}$$

For matrix $\boldsymbol{X}$, we extract these as

$$d_{\boldsymbol{V}} f(\boldsymbol{X}) = \operatorname{vec}(\nabla f(\boldsymbol{X}))^t \operatorname{vec}(\boldsymbol{V}) = \operatorname{tr}((\nabla f(\boldsymbol{X}))^t \boldsymbol{V})$$
$$d^2_{\boldsymbol{V}} f(\boldsymbol{X}) = \operatorname{vec}(\boldsymbol{V})^t \boldsymbol{H} \operatorname{vec}(\boldsymbol{V})$$

Proof.

The gradients are covered in Prop 3.2.1 of MM Optimization. For matrix gradients, the trace function induces an inner product on matrices by analogy with the ordinary inner product for vectors. The formula $\operatorname{tr}(\boldsymbol{A}^t \boldsymbol{B}) = \operatorname{vec}(\boldsymbol{A})^t \operatorname{vec}(\boldsymbol{B})$ converts the trace inner product to the vector inner product. □

Warm up 1: find $d_{\boldsymbol{V}}[f(\boldsymbol{X})^{-1}]$ for $\boldsymbol{X}$ and $f(\boldsymbol{X})$ matrices.

$$
\begin{aligned}
d_{\boldsymbol{V}} f(\boldsymbol{X})^{-1} &= \lim_{\substack{h \to 0 \\ \boldsymbol{W} \to \boldsymbol{V}}} \frac{f(\boldsymbol{X} + t\boldsymbol{W})^{-1} - f(\boldsymbol{X})^{-1}}{t} \\
&= \lim_{\substack{h \to 0 \\ \boldsymbol{W} \to \boldsymbol{V}}} f(\boldsymbol{X} + t\boldsymbol{W})^{-1} \frac{1 - f(\boldsymbol{X} + t\boldsymbol{W}) f(\boldsymbol{X})^{-1}}{t} \\
&= \lim_{\substack{h \to 0 \\ \boldsymbol{W} \to \boldsymbol{V}}} f(\boldsymbol{X} + t\boldsymbol{W})^{-1} \frac{f(\boldsymbol{X}) - f(\boldsymbol{X} + t\boldsymbol{W})}{t} f(\boldsymbol{X})^{-1} \\
&= -f(\boldsymbol{X})^{-1} d_{\boldsymbol{V}} f(\boldsymbol{X}) f(\boldsymbol{X})^{-1}.
\end{aligned}
$$

Warm up 2: directional derivative of log determinants

For a symmetric positive definite $\boldsymbol{X}$ and a symmetric direction $\boldsymbol{V}$,

$$
\begin{aligned}
d_{\boldsymbol{V}} \ln\det(\boldsymbol{X}) &= \lim_{\substack{h\to 0 \\ \boldsymbol{W}\to\boldsymbol{V}}} \frac{\ln\det(\boldsymbol{X}+t\boldsymbol{W}) - \ln\det(\boldsymbol{X})}{t} \\
&= \lim_{\substack{h\to 0 \\ \boldsymbol{W}\to\boldsymbol{V}}} \frac{\ln\left[\det(\boldsymbol{X}^{1/2})\det(\boldsymbol{I}+t\boldsymbol{X}^{-1/2}\boldsymbol{W}\boldsymbol{X}^{-1/2})\boldsymbol{X}^{1/2}\right] - \ln\det(\boldsymbol{X})}{t} \\
&= \lim_{\substack{h\to 0 \\ \boldsymbol{W}\to\boldsymbol{V}}} \frac{\ln\det(\boldsymbol{I}+t\boldsymbol{X}^{-1/2}\boldsymbol{W}\boldsymbol{X}^{-1/2})}{t} \\
&= \lim_{\substack{h\to 0 \\ \boldsymbol{W}\to\boldsymbol{V}}} \frac{\sum \ln(1+t\lambda_i)}{t} \quad \left(\lambda = \text{eigenvalues of } \boldsymbol{X}^{-1/2}\boldsymbol{W}\boldsymbol{X}^{-1/2}\right) \\
&\approx \lim_{\substack{h\to 0 \\ \boldsymbol{W}\to\boldsymbol{V}}} \frac{\sum t\lambda_i}{t} \quad (\ln(1+x)\approx x, \text{ for small } x) \\
&= \sum \lambda_i = \operatorname{tr}\left(\boldsymbol{X}^{-1/2}\boldsymbol{V}\boldsymbol{X}^{-1/2}\right) = \operatorname{tr}(\boldsymbol{X}^{-1}\boldsymbol{V}).
\end{aligned}
$$

Detailed example: evaluate gradient of LMM

The loglikelihood for the ith sample in a linear mixed model is

$$\ell_i(\boldsymbol{\beta}, \boldsymbol{L}, \sigma^2) = -\frac{1}{2} \log \det \Omega_i - \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}_i\boldsymbol{\beta})^T \Omega_i^{-1} (\boldsymbol{y} - \boldsymbol{X}_i\boldsymbol{\beta}).$$

Prove that the gradient is

$$\begin{aligned}
\nabla_{\boldsymbol{\beta}} \ell_i(\boldsymbol{\beta}, \boldsymbol{L}, \sigma^2) &= \boldsymbol{X}_i^T \Omega_i^{-1} \boldsymbol{r}_i, \\
\nabla_{\sigma^2} \ell_i(\boldsymbol{\beta}, \boldsymbol{L}, \sigma^2) &= -\frac{1}{2} \operatorname{tr}(\Omega_i^{-1}) + \frac{1}{2} \boldsymbol{r}_i^T \Omega_i^{-2} \boldsymbol{r}_i, \\
\nabla_{\boldsymbol{L}} \ell_i(\boldsymbol{\beta}, \boldsymbol{L}, \sigma^2) &= -\boldsymbol{Z}_i^T \Omega_i^{-1} \boldsymbol{Z}_i \boldsymbol{L} + \boldsymbol{Z}_i^T \Omega_i^{-1} \boldsymbol{r}_i \boldsymbol{r}_i^T \Omega_i^{-1} \boldsymbol{Z}_i \boldsymbol{L},
\end{aligned}$$

where $\Omega_i = \sigma^2 \boldsymbol{I} + \boldsymbol{Z}_i \Sigma \boldsymbol{Z}_i^t = \sigma^2 \boldsymbol{I} + \boldsymbol{Z}_i \boldsymbol{L}\boldsymbol{L}^t \boldsymbol{Z}_i^t$ and $\boldsymbol{r} = \boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}$. We optimize over the Cholesky factor $\boldsymbol{L}$ instead of Σ directly because the later needs to be positive semidefinite. The diagonal entries of $\boldsymbol{L}$ must be positive, simpler constraint.

1st term: $d_{\boldsymbol{v}}(\boldsymbol{y}-\boldsymbol{X}_i\boldsymbol{\beta})^T\Omega_i^{-1}(\boldsymbol{y}-\boldsymbol{X}_i\boldsymbol{\beta})$ wrt to $\boldsymbol{\beta}$

Using symmetry of Ω_i,

$$
\begin{aligned}
&-\frac{1}{2}d_{\boldsymbol{v}}(\boldsymbol{y}-\boldsymbol{X}_i\boldsymbol{\beta})^T\Omega_i^{-1}(\boldsymbol{y}-\boldsymbol{X}_i\boldsymbol{\beta}) \\
=&-\frac{1}{2}\lim_{\substack{t\to 0\\ \boldsymbol{w}\to\boldsymbol{v}}}\frac{(\boldsymbol{y}-\boldsymbol{X}_i(\boldsymbol{\beta}+t\boldsymbol{w}))^t\Omega_i^{-1}(\boldsymbol{y}-\boldsymbol{X}_i(\boldsymbol{\beta}+t\boldsymbol{w}))-(\boldsymbol{y}-\boldsymbol{X}_i\boldsymbol{\beta})^T\Omega_i^{-1}(\boldsymbol{y}-\boldsymbol{X}}{t} \\
=&-\frac{1}{2}\lim_{\substack{t\to 0\\ \boldsymbol{w}\to\boldsymbol{v}}}\frac{-t(\boldsymbol{y}-\boldsymbol{X}_i\boldsymbol{\beta})^t\Omega_i^{-1}\boldsymbol{X}_i\boldsymbol{w}-t(\boldsymbol{X}_i\boldsymbol{w})^t\Omega_i^{-1}(\boldsymbol{y}-\boldsymbol{X}_i\boldsymbol{\beta})+O(t^2)}{t} \\
=&(\boldsymbol{y}-\boldsymbol{X}_i\boldsymbol{\beta})^t\Omega_i^{-1}\boldsymbol{X}_i\boldsymbol{v}.
\end{aligned}
$$

To extract gradients, we set

$$(\nabla_{\boldsymbol{\beta}}\ell)^t\boldsymbol{v}\equiv(\boldsymbol{y}-\boldsymbol{X}_i\boldsymbol{\beta})^t\Omega_i^{-1}\boldsymbol{X}_i\boldsymbol{v}$$

So as desired,

$$\nabla_{\boldsymbol{\beta}}\ell=\boldsymbol{X}_i^t\Omega_i^{-1}(\boldsymbol{y}-\boldsymbol{X}_i\boldsymbol{\beta}).$$

2nd term part I: $d_v \ln\det(\sigma^2 \boldsymbol{I} + \boldsymbol{Z}\boldsymbol{L}\boldsymbol{L}^t\boldsymbol{Z}^t)$ wrt to σ^2

Chain rule says we must compute

$$d_{d_v\sigma^2 \boldsymbol{I}} \ln\det(\sigma^2 \boldsymbol{I} + \boldsymbol{Z}\boldsymbol{L}\boldsymbol{L}^t\boldsymbol{Z}^t).$$

First evaluate the direction

$$d_v\sigma^2 \boldsymbol{I} = \lim_{\substack{t\to 0\\ w\to v}} \frac{(\sigma^2 + tw)\boldsymbol{I} - \sigma^2 \boldsymbol{I}}{t} = \lim_{\substack{t\to 0\\ w\to v}} \frac{tw\boldsymbol{I}}{t} = v\boldsymbol{I}$$

Thus

$$\begin{aligned} & -\frac{1}{2} d_{d_v\sigma^2 \boldsymbol{I}} \ln\det(\sigma^2 \boldsymbol{I} + \boldsymbol{Z}\boldsymbol{L}\boldsymbol{L}^t\boldsymbol{Z}^t) \\ = & -\frac{1}{2} d_{v\boldsymbol{I}} \ln\det(\sigma^2 \boldsymbol{I} + \boldsymbol{Z}\boldsymbol{L}\boldsymbol{L}^t\boldsymbol{Z}^t) \\ = & -\frac{1}{2} \operatorname{tr}\left((\sigma^2 \boldsymbol{I} + \boldsymbol{Z}\boldsymbol{L}\boldsymbol{L}^t\boldsymbol{Z}^t)^{-1} v\boldsymbol{I}\right) \quad \text{(warm up 2)} \end{aligned}$$

2nd term part II: $d_v \boldsymbol{r}^t(\sigma^2 \boldsymbol{I} + \boldsymbol{Z}\boldsymbol{L}\boldsymbol{L}^t\boldsymbol{Z}^t)^{-1}\boldsymbol{r}$ wrt to σ^2

In our notation $\Omega_i = \sigma^2 \boldsymbol{I} + \boldsymbol{Z}\boldsymbol{L}\boldsymbol{L}^t\boldsymbol{Z}^t$, so

$$
\begin{aligned}
& d_v \left[-\frac{1}{2} \boldsymbol{r}^t (\sigma^2 \boldsymbol{I} + \boldsymbol{Z}\boldsymbol{L}\boldsymbol{L}^t\boldsymbol{Z}^t)^{-1} \boldsymbol{r} \right] \\
= & -\frac{1}{2} \boldsymbol{r}^t d_v \left[(\sigma^2 \boldsymbol{I} + \boldsymbol{Z}\boldsymbol{L}\boldsymbol{L}^t\boldsymbol{Z}^t)^{-1} \right] \boldsymbol{r} \\
= & \frac{1}{2} \boldsymbol{r}^t \Omega_i^{-1} d_v (\sigma^2 \boldsymbol{I} + \boldsymbol{Z}\boldsymbol{L}\boldsymbol{L}^t\boldsymbol{Z}^t) \Omega_i^{-1} \boldsymbol{r} \qquad \text{(warm up 1)} \\
= & \frac{1}{2} \boldsymbol{r}^t \Omega_i^{-1} \lim_{\substack{t \to 0 \\ w \to v}} \frac{(\sigma^2 + tw)\boldsymbol{I} + \boldsymbol{Z}\boldsymbol{L}\boldsymbol{L}^t\boldsymbol{Z}^t - (\sigma^2 \boldsymbol{I} + \boldsymbol{Z}\boldsymbol{L}\boldsymbol{L}^t\boldsymbol{Z}^t)}{t} \Omega_i^{-1} \boldsymbol{r} \\
= & \frac{1}{2} \boldsymbol{r} \Omega_i^{-1} \lim_{\substack{t \to 0 \\ w \to v}} \frac{tw\boldsymbol{I}}{t} \Omega_i^{-1} \boldsymbol{r} \\
= & \frac{1}{2} \boldsymbol{r}^t \Omega_i^{-1} v \boldsymbol{I} \Omega_i^{-1} \boldsymbol{r} \\
= & \frac{1}{2} \boldsymbol{r}^t \Omega_i^{-2} \boldsymbol{r} v \quad (v \text{ is constant})
\end{aligned}
$$

2rd term: extracting gradients

In summary, the full 1st directional derivative wrt to σ^2 is

$$d_v \ell = -\frac{1}{2} \operatorname{tr}\left((\sigma^2 \boldsymbol{I} + \boldsymbol{Z}\boldsymbol{L}\boldsymbol{L}^t\boldsymbol{Z}^t)^{-1} v \boldsymbol{I}\right) + \frac{1}{2}\boldsymbol{r}^t\Omega_i^{-2}\boldsymbol{r}v$$

To extract gradients, we set

$$v(\nabla_{\sigma^2}\ell) \equiv -\frac{1}{2} \operatorname{tr}\left((\sigma^2 \boldsymbol{I} + \boldsymbol{Z}\boldsymbol{L}\boldsymbol{L}^t\boldsymbol{Z}^t)^{-1}\right) v + \frac{1}{2}\boldsymbol{r}^t\Omega_i^{-2}\boldsymbol{r}v$$

So as desired,

$$\nabla_{\sigma^2}\ell = -\frac{1}{2} \operatorname{tr}\left(\Omega_i^{-1}\right) + \frac{1}{2}\boldsymbol{r}^t\Omega_i^{-2}\boldsymbol{r}$$

3rd term part I: $d_{\boldsymbol{V}} \ln\det(\sigma^2 \boldsymbol{I} + \boldsymbol{Z}\boldsymbol{L}\boldsymbol{L}^t\boldsymbol{Z}^t)$ wrt to $\boldsymbol{L}$

Chain rule says we must compute

$$d_{d_{\boldsymbol{V}}\boldsymbol{Z}\boldsymbol{L}\boldsymbol{L}^t\boldsymbol{Z}^t} \ln\det(\sigma^2 \boldsymbol{I} + \boldsymbol{Z}\boldsymbol{L}\boldsymbol{L}^t\boldsymbol{Z}^t).$$

First evaluate the direction

$$\begin{aligned} d_{\boldsymbol{V}}\boldsymbol{Z}\boldsymbol{L}\boldsymbol{L}^t\boldsymbol{Z}^t &= \lim_{\substack{t\to 0 \\ \boldsymbol{W}\to\boldsymbol{V}}} \frac{\boldsymbol{Z}(\boldsymbol{L}+t\boldsymbol{W})(\boldsymbol{L}+t\boldsymbol{W})^t\boldsymbol{Z}^t - \boldsymbol{Z}\boldsymbol{L}\boldsymbol{L}^t\boldsymbol{Z}^t}{t} \\ &= \lim_{\substack{t\to 0 \\ \boldsymbol{W}\to\boldsymbol{V}}} \frac{t\boldsymbol{Z}\boldsymbol{L}\boldsymbol{W}^t\boldsymbol{Z}^t + t\boldsymbol{Z}\boldsymbol{W}\boldsymbol{L}^t\boldsymbol{Z}^t + t^2\boldsymbol{Z}\boldsymbol{W}\boldsymbol{W}^t\boldsymbol{Z}^t}{t} \\ &= \boldsymbol{Z}\boldsymbol{L}\boldsymbol{V}^t\boldsymbol{Z}^t + \boldsymbol{Z}\boldsymbol{V}\boldsymbol{L}^t\boldsymbol{Z}^t = 2\boldsymbol{Z}\boldsymbol{V}\boldsymbol{L}^t\boldsymbol{Z}^t. \end{aligned}$$

Thus

$$\begin{aligned} & d_{\boldsymbol{V}} \ln\det(\sigma^2 \boldsymbol{I} + \boldsymbol{Z}\boldsymbol{L}\boldsymbol{L}^t\boldsymbol{Z}) \\ =& d_{2\boldsymbol{Z}\boldsymbol{V}\boldsymbol{L}^t\boldsymbol{Z}^t} \ln\det(\sigma^2 \boldsymbol{I} + \boldsymbol{Z}\boldsymbol{L}\boldsymbol{L}^t\boldsymbol{Z}) && \text{(chain rule)} \\ =& 2\,\mathrm{tr}\left((\sigma^2 \boldsymbol{I} + \boldsymbol{Z}\boldsymbol{L}\boldsymbol{L}^t\boldsymbol{Z}^t)^{-1}(\boldsymbol{Z}\boldsymbol{V}\boldsymbol{L}^t\boldsymbol{Z}^t)\right) && \text{(warm up 2)} \\ =& 2\,\mathrm{tr}\left([\boldsymbol{L}^t\boldsymbol{Z}^t(\sigma^2 \boldsymbol{I} + \boldsymbol{Z}\boldsymbol{L}\boldsymbol{L}^t\boldsymbol{Z}^t)^{-1}\boldsymbol{Z}]\boldsymbol{V}\right) \end{aligned}$$

3rd term part II: $d_{\boldsymbol{V}} \boldsymbol{r}^t(\sigma^2 \boldsymbol{I} + \boldsymbol{Z}\boldsymbol{L}\boldsymbol{L}^t\boldsymbol{Z})^{-1}\boldsymbol{r}$ wrt $\boldsymbol{L}$

In our notation $\Omega = \sigma^2 \boldsymbol{I} + \boldsymbol{Z}\boldsymbol{L}\boldsymbol{L}^t\boldsymbol{Z}$, so

$$
\begin{aligned}
& d_{\boldsymbol{V}} \boldsymbol{r}^t(\sigma^2 \boldsymbol{I} + \boldsymbol{Z}\boldsymbol{L}\boldsymbol{L}^t\boldsymbol{Z})^{-1}\boldsymbol{r} \\
=& \boldsymbol{r}^t d_{\boldsymbol{V}} \left[(\sigma^2 \boldsymbol{I} + \boldsymbol{Z}\boldsymbol{L}\boldsymbol{L}^t\boldsymbol{Z})^{-1}\right] \boldsymbol{r} \\
=& -\boldsymbol{r}^t \Omega_i^{-1} d_{\boldsymbol{V}} \left[(\sigma^2 \boldsymbol{I} + \boldsymbol{Z}\boldsymbol{L}\boldsymbol{L}^t\boldsymbol{Z})\right] \Omega_i^{-1}\boldsymbol{r} \qquad \text{(warm up 1)} \\
=& -\boldsymbol{r}^t \Omega_i^{-1} \left[\lim_{\substack{t\to 0 \\ \boldsymbol{W}\to\boldsymbol{V}}} \frac{\sigma^2 \boldsymbol{I} + \boldsymbol{Z}(\boldsymbol{L}+t\boldsymbol{W})(\boldsymbol{L}+t\boldsymbol{W})^t\boldsymbol{Z}^t - \sigma^2 \boldsymbol{I} - \boldsymbol{Z}\boldsymbol{L}\boldsymbol{L}^t\boldsymbol{Z}^t}{t} \right] \Omega_i^{-1}\boldsymbol{r} \\
=& -\boldsymbol{r}^t \Omega_i^{-1} \left[\boldsymbol{Z}\boldsymbol{L}\boldsymbol{V}\boldsymbol{Z}^t + \boldsymbol{Z}\boldsymbol{V}\boldsymbol{L}^t\boldsymbol{Z}^t\right] \Omega_i^{-1}\boldsymbol{r} \\
=& -2\,\mathrm{tr}\left(\boldsymbol{r}^t \Omega_i^{-1}\boldsymbol{Z}\boldsymbol{V}\boldsymbol{L}^t\boldsymbol{Z}^t\Omega_i^{-1}\boldsymbol{r}\right) \\
=& -2\,\mathrm{tr}\left(\boldsymbol{L}^t\boldsymbol{Z}^t\Omega_i^{-1}\boldsymbol{r}\boldsymbol{r}^t\Omega_i^{-1}\boldsymbol{Z}\boldsymbol{V}\right)
\end{aligned}
$$

3rd term: extracting gradients

To extract gradients, we set

$$\begin{aligned}\operatorname{tr}\left((\nabla_{\boldsymbol{L}}\ell)^t \boldsymbol{V}\right) \quad &\equiv \quad -\operatorname{tr}\left([\boldsymbol{L}^t \boldsymbol{Z}^t(\sigma^2 \boldsymbol{I} + \boldsymbol{Z}\boldsymbol{L}\boldsymbol{L}^t\boldsymbol{Z}^t)^{-1}\boldsymbol{Z}]\boldsymbol{V}\right) \\ &\qquad +\operatorname{tr}\left([\boldsymbol{L}^t \boldsymbol{Z}^t \Omega_i^{-1}\boldsymbol{r}\boldsymbol{r}^t\Omega_i^{-1}\boldsymbol{Z}]\boldsymbol{V}\right)\end{aligned}$$

So as desired,

$$\nabla_{\boldsymbol{L}}\ell = -\boldsymbol{Z}^t(\sigma^2 \boldsymbol{I} + \boldsymbol{Z}\boldsymbol{L}\boldsymbol{L}^t\boldsymbol{Z}^t)^{-1}\boldsymbol{Z}\boldsymbol{L} + \boldsymbol{Z}^t\Omega_i^{-1}\boldsymbol{r}\boldsymbol{r}^t\Omega_i^{-1}\boldsymbol{Z}\boldsymbol{L}.$$

Learning resources

1. Chapter 3 of MM Optimization by Kenneth Lange
2. Chapter 3 of Introduction to optimization and semidifferential calculus by Michel Delfour
3. Chapters 5, 6, 8, 9 of Matrix differential calculus with applications in statistics and econometrics by Magnus and Neudecker