Matrix Differential Calculus

Zhang Le

Centre for Speech Technology Research
University of Edinburgh

October 23, 2006

What's the hell of this?

Matrix Differential Calcus

- Straightforward extension to scalar chain rule?

What's the hell of this?

Matrix Differential Calcus

- Straightforward extension to scalar chain rule?
- Complex subject of its own?

What's the hell of this?

Matrix Differential Calcus

- Straightforward extension to scalar chain rule?
- Complex subject of its own?
- Or, can be something in between?

Main references

- *Matrix Differential Calculus with Applications in Statistics and Econometrics*, 2nd Ed (Magnus and Neudecker 1999), QA188 Mag, JCMB
- *Old and New Matrix Algebra Useful for Statistics* (Minka 2000) http://research.microsoft.com/~minka/papers/matrix/

Today's talk:

- 1st-order only, sorry no Hessian!
- Only discuss $f : R^n \rightarrow R$

1D case

- From text book we have:

$$\lim_{\mu \to 0} \frac{\phi(c+\mu) - \phi(c)}{\mu} = \phi'(c)$$

1D case

- From text book we have:

$$\lim_{\mu \to 0} \frac{\phi(c+\mu) - \phi(c)}{\mu} = \phi'(c)$$

- or,

$$\phi(c+\mu) = \phi(c) + \phi'(c)\mu + r_c(\mu)$$

1D case (cont.)

Define differential of ϕ at c with increment μ:

$$d\phi(c;\mu) = \phi'(c)\mu$$

or $d\phi(c;\mu)$ is part of $\phi(c+\mu) - \phi(c)$ which is **linear** in μ

Geometric interpretation

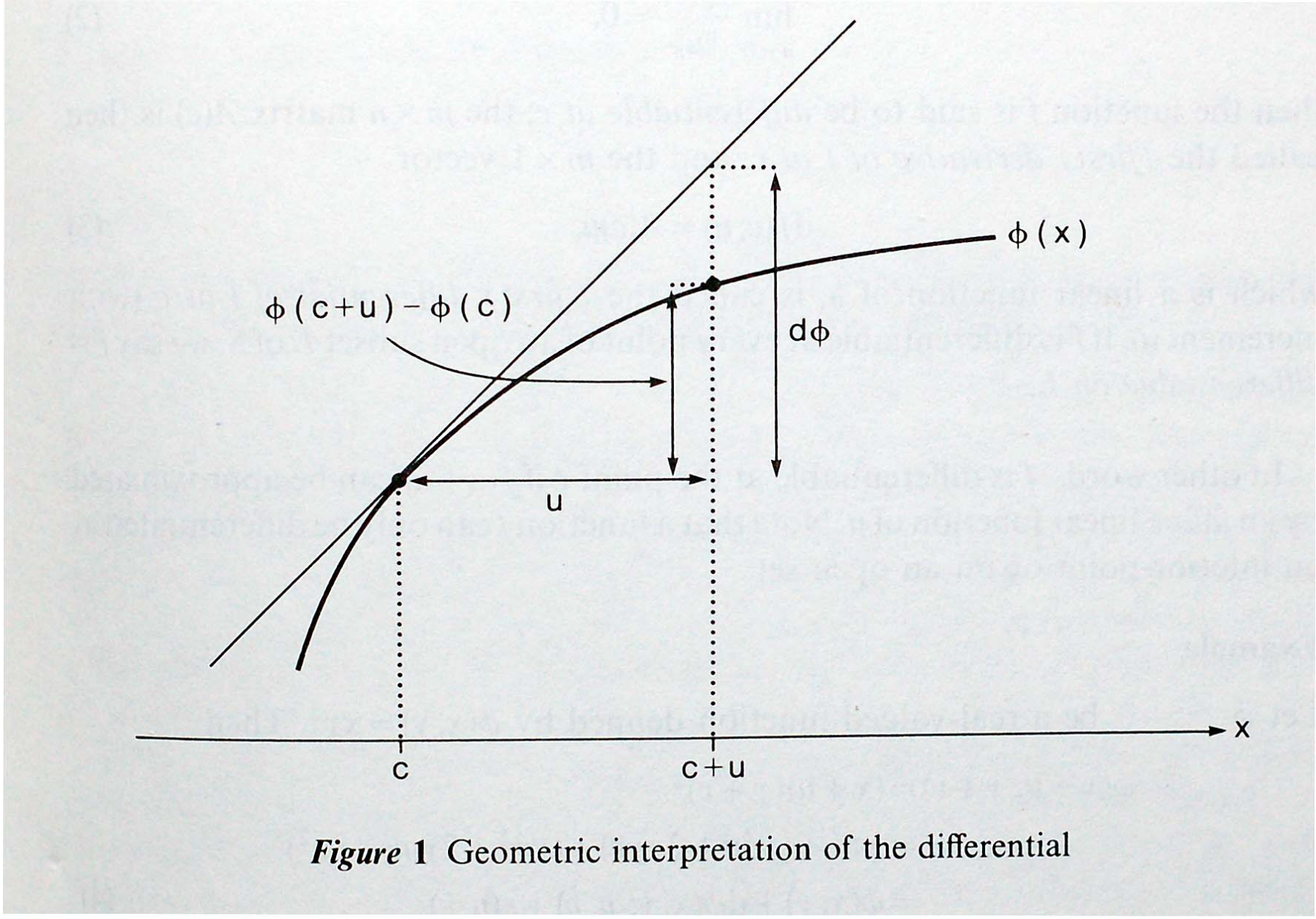

***Figure* 1** Geometric interpretation of the differential

Open questions

1. Differential
2. Derivative
3. Gradient

Are they the same?

Vector case

$$f(c+\mu) = f(c) + A(c)\mu + R_c(\mu)$$

- c, μ: $n \times 1$ vector
- $A(c)$: $1 \times n$ row vector, also called Jacobian
- $A^T(c)$:$n \times 1$ vector, also called the gradient
- $R_c(\mu)$: higher-order term so that $R_c(\mu) \to 0$ when $\mu \to 0$

Question answered

1. Differential: $A(c)\mu$, a real number (nothing to do with infinitly small number...)
2. Derivative: $A(c)$, a row vector $[\frac{\partial f(c)}{\partial c_i}, \dots]$
3. Gradient: $A^T(c)$, transpose of derivative

So **Derivative vector** is just the coefficient of **Differential.** (for $f : R^n \to R^m$, derivative will be an $m \times n$ matrix (Jacobian matrix)

Matrix differential rules

1.

$$
\begin{aligned}
dA &= 0 \\
d(\alpha X) &= \alpha dX \\
d(X+Y) &= dX + dY \\
d(tr(X)) &= tr(dX) \\
d(XY) &= (dX)Y + XdY \\
dX^{-1} &= -X^{-1}(dX)X^{-1} \\
d|X| &= |X| tr(X^{-1}dX) \\
d\log|X| &= tr(X^{-1}dX)
\end{aligned}
$$

Matrix differential rules

1.

$$
\begin{aligned}
dA &= 0 \\
d(\alpha X) &= \alpha dX \\
d(X+Y) &= dX + dY \\
d(tr(X)) &= tr(dX) \\
d(XY) &= (dX)Y + XdY \\
dX^{-1} &= -X^{-1}(dX)X^{-1} \\
d|X| &= |X| tr(X^{-1}dX) \\
d\log|X| &= tr(X^{-1}dX)
\end{aligned}
$$

2. But be careful here:

$$
\begin{aligned}
d(AXB) &= Ad(X)B = BAdX \\
d(AX^TB) &= Ad(X^T)B = A^TB^TdX
\end{aligned}
$$

Compute the derivative/gradient

1. Compute the differential using matrix chain rules
2. Message the result into canonical form (dX is at the rightmost)
3. Read $A(c)$ off as coefficient of dX as **derivative vector**
4. (optionally) $A^T(c)$ is the **gradient vector**

Example: Gaussian differential

$$\log p(\mathbf{x}) = -\frac{1}{2}\left\{(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu) + d\log 2\pi + \log|\Sigma|\right\}$$

Let's compute the differential of $\log p(\mathbf{x})$ w.r.t μ and Σ step by step

Differential w.r.t. μ

$$
\begin{aligned}
d \log p(\mathbf{x}) &= -\frac{1}{2}[(d(\mathbf{x}-\mu)^T \Sigma^{-1}(\mathbf{x}-\mu) + (\mathbf{x}-\mu)^T \Sigma^{-1} d(\mathbf{x}-\mu)] \\
&= -\frac{1}{2}[2(\mathbf{x}-\mu)^T \Sigma^{-1} d(\mathbf{x}-\mu)] \\
&= (\mathbf{x}-\mu)^T \Sigma^{-1} d\mu
\end{aligned}
$$

1. Derivative: $(\mathbf{x}-\mu)^T \Sigma^{-1}$
2. Gradient: $\Sigma^{-1}(\mathbf{x}-\mu)$
3. diag. cov. case: $(x-\mu)/\sigma^2$

Differential w.r.t. Σ

$$
\begin{aligned}
d \log p(\mathbf{x}) &= -\frac{1}{2}[(\mathbf{x}-\mu)^T(d\Sigma^{-1})(\mathbf{x}-\mu) + tr(\Sigma^{-1}d\Sigma)] \\
&= -\frac{1}{2}[(\mathbf{x}-\mu)(\mathbf{x}-\mu)^T(-\Sigma^{-1}(d\Sigma)\Sigma^{-1}) + tr(\Sigma^{-1}d\Sigma)] \\
&= -\frac{1}{2}[-\Sigma^{-1}(\mathbf{x}-\mu)(\mathbf{x}-\mu)^T\Sigma^{-1}d\Sigma + tr(\Sigma^{-1}d\Sigma)]
\end{aligned}
$$

1. Derivative/Gradient: $-\frac{1}{2}[\Sigma^{-1} - \Sigma^{-1}(\mathbf{x}-\mu)(\mathbf{x}-\mu)^T\Sigma^{-1}]$
2. diag. cov. case: $-\frac{1}{2\sigma^2}(1 - \frac{(x-\mu)^2}{\sigma^2})$

Recap

- Differential (R), derivative ($1 \times n$) & gradient ($n \times 1$)
- Steps to compute derivative:
  - Compute differential
  - Message result into canonical form
  - Read off coefficient of dX