Gradient Calculations for the Mean, Covariance Matrix Parameterization of the Multivariate Normal

Jason D. M. Rennie
jrennie@gmail.com

August 18, 2006

1 Introduction

We use $\boldsymbol{\mu} \in \mathbb{R}^d$ to parameterize the mean vector and $\Lambda \in \mathbb{R}^{d \times k}$ to parameterize the covariance matrix. The covariance matrix is $S = \Lambda\Lambda^T$. Note that S is symmetric ($S = S^T$) and positive semi-definite ($\mathbf{x}S\mathbf{x}^T \geq 0$. The likelihood of a set of n data points, $X \in \mathbb{R}^{n \times d}$ is

$$P(X|\boldsymbol{\mu}, \Lambda) = \frac{1}{(2\pi)^{nd/2}|S|^{n/2}} \exp\left(-\frac{1}{2}\sum_i (X_{i\cdot} - \boldsymbol{\mu})S^{-1}(X_{i\cdot} - \boldsymbol{\mu})^T\right). \quad (1)$$

Maximum likelihood parameters are those that maximize the likelihood, or, equivalently, minimize the negative log-likelihood,

$$J = P(X|\boldsymbol{\mu}, \Lambda) = C + \frac{1}{2}\left(n \log|S| + \sum_{a=1}^{n}(X_a - \boldsymbol{\mu})S^{-1}(X_a - \boldsymbol{\mu})^T\right), \quad (2)$$

where C is a constant (not a function of $\boldsymbol{\mu}$, or Λ). One option for learning of the parameters is gradient descent[1]. It is well known that negative log-likelihood is not convex in this traditional (mean, covariance matrix) parameterization. However, we can use gradient descent to find a local minimum.

[1] By "gradient descent," we are not referring specifically to the simple algorithm which uses the negative gradient as its direction at each step, but rather the family of algoritms which use only (current and past) gradient calculations to choose a direction at each step.

2 Setup

To simplify the full gradient calculation, we break the objective into two its two important terms, $J = C + \frac{1}{2}(J_1 + J_2)$, where

$$J_1 = n \log |S| \tag{3}$$

$$J_2 = \sum_a (X_a - \boldsymbol{\mu}) S^{-1} (X_a - \boldsymbol{\mu})^T = \sum_{a,b,c} [X_a - \boldsymbol{\mu}]_b S^{-1}_{bc} [X_a - \boldsymbol{\mu}]_c. \tag{4}$$

We provide a number of intermediate partial derivative calculations:

- $\frac{\partial S_{kl}}{\partial \Lambda_{ij}} = \frac{\partial (\Lambda\Lambda^T)_{kl}}{\partial \Lambda_{ij}} = \frac{\partial (\sum_a \Lambda_{ka}\Lambda_{la})}{\partial \Lambda_{ij}} = \delta_{k=i}\Lambda_{lj} + \Lambda_{kj}\delta_{l=i}$
- $\frac{\partial a^T M a}{\partial M_{ij}} = \sum_{k,l} a_k \frac{\partial M_{kl}}{\partial M_{ij}} a_l = a_i a_j$
- $\frac{\partial (S^{-1})_{kl}}{\partial S_{ij}} = -\left(S^{-1} \frac{\partial S}{\partial S_{ij}} S^{-1}\right)_{kl} = -S^{-1}_{ki} S^{-1}_{jl}$ (page 8 of [1])
  Note: $\frac{\partial (ABA)_{kl}}{\partial B_{ij}} = \frac{\partial \left(\sum_{a,b} A_{ka} B_{ab} A_{bl}\right)}{\partial B_{ij}} = A_{ki} A_{jl}$
- $\frac{\partial \log |S|}{\partial S_{ij}} = S^{-1}_{ji}$ (page 7 of [1], eqn. 10)
  Note: $\partial \log |S| = \mathbf{Tr}(S^{-1} \partial S) = \sum_{i,j} S^{-1}_{ji} \partial S_{ij}$

3 Gradient Calculations

Partial derivative of J_1 wrt Λ:

$$\frac{\partial J_1}{\partial \Lambda_{ij}} = n \sum_{k,l} \frac{\partial \log |S|}{\partial S_{kl}} \frac{\partial S_{kl}}{\partial \Lambda_{ij}} = n \sum_{k,l} S^{-1}_{lk} (\delta_{k=i}\Lambda_{lj} + \Lambda_{kj}\delta_{l=i}) \tag{5}$$

$$= n \left[\sum_l S^{-1}_{il} \Lambda_{lj} + \sum_k S^{-1}_{ik} \Lambda_{kj} \right] = 2n S^{-1}_{i\cdot} \Lambda_{\cdot j} \tag{6}$$

$$\frac{\partial J_1}{\partial \Lambda} = 2n S^{-1} \Lambda \tag{7}$$

Partial derivative of J_2 wrt $\boldsymbol{\mu}$:

$$\frac{\partial J_2}{\partial \mu_j} = -\sum_{a,c} S^{-1}_{jc} [X_a - \boldsymbol{\mu}]_c - \sum_{a,b} [X_a - \boldsymbol{\mu}]_b S^{-1}_{bj} = -2 \sum_{a,b} [X_a - \boldsymbol{\mu}]_b S^{-1}_{bj} \tag{8}$$

$$\frac{\partial J_2}{\partial \boldsymbol{\mu}} = -2 \sum_a [X_a - \boldsymbol{\mu}] S^{-1}_{\cdot j} \tag{9}$$

Partial derivative of J_2 wrt Λ:

$$
\begin{aligned}
\frac{\partial J_2}{\partial \Lambda_{ij}} &= \sum_{a,b,c,k,l} [X_a - \boldsymbol{\mu}]_b [X_a - \boldsymbol{\mu}]_c \frac{\partial S^{-1}_{bc}}{\partial S_{kl}} \frac{\partial S_{kl}}{\partial \Lambda_{ij}} && (10)\\
&= -\sum_{a,b,c,k,l} [X_a - \boldsymbol{\mu}]_b [X_a - \boldsymbol{\mu}]_c S^{-1}_{bk} S^{-1}_{lc} (\delta_{k=i} \Lambda_{lj} + \Lambda_{kj} \delta_{l=i}) && (11)\\
&= -\sum_{a,b,c,l} [X_a - \boldsymbol{\mu}]_b [X_a - \boldsymbol{\mu}]_c S^{-1}_{bi} S^{-1}_{lc} \Lambda_{lj} - \sum_{a,b,c,k} [X_a - \boldsymbol{\mu}]_b [X_a - \boldsymbol{\mu}]_c S^{-1}_{bk} S^{-1}_{ic} \Lambda_{kj}\\
&= -\sum_{a,b,c,l} S^{-1}_{ib} [X_a - \boldsymbol{\mu}]_b [X_a - \boldsymbol{\mu}]_c S^{-1}_{cl} \Lambda_{lj} - \sum_{a,b,c,k} S^{-1}_{ic} [X_a - \boldsymbol{\mu}]_c [X_a - \boldsymbol{\mu}]_b S^{-1}_{bk} \Lambda_{kj}\\
&= -2 \sum_{a} S^{-1}_{i\cdot} [X_a - \boldsymbol{\mu}]^T [X_a - \boldsymbol{\mu}] S^{-1} \Lambda_{\cdot j} && (12)\\
&= -2 S^{-1}_{i\cdot} [X - \boldsymbol{\mu}]^T [X - \boldsymbol{\mu}] S^{-1} \Lambda_{\cdot j} && (13)\\
\frac{\partial J_2}{\partial \Lambda} &= -2 S^{-1} [X - \boldsymbol{\mu}]^T [X - \boldsymbol{\mu}] S^{-1} \Lambda && (14)
\end{aligned}
$$

Notation

a, α	scalar
$\mathbf{a}$, $\boldsymbol{\alpha}$	(row) vector
A	matrix
$\mathrm{diag}(A)$	diagonal of A, treated as a row vector
$\mathrm{diag}(\boldsymbol{\alpha})$	diagonal matrix, with diagonal elements taken from $\boldsymbol{\alpha}$
A_{ab}	scalar from a^{th} row, b^{th} column of A
A^{-1}_{ab}	scalar from a^{th} row, b^{th} column of A^{-1}
$A_{\cdot b}$	b^{th} column of A (as a column vector)
$A_{a\cdot}$	a^{th} row of A (as a row vector)
AB	matrix multiplication
$\mathbf{a}\mathbf{b}^T$	vector product
$A * B$, $\mathbf{a} * \mathbf{b}$	element-wise multiplication
A/B, $\mathbf{a}/\mathbf{b}$	element-wise division
$A + \mathbf{a}$, $A - \mathbf{a}$	add/subtract $\mathbf{a}$ from each row of A
$A * \mathbf{a}$, $\frac{A}{\mathbf{a}}$	multiply/divide each row of A by $\mathbf{a}$

References

[1] K. B. Petersen and M. S. Pedersen. The matrix cookbook. http://matrixcookbook.com, February 2006.