Gaussian maximum likelihood estimation: two useful facts

Peter Bartlett. April 3, 2007.

These notes present two simple but useful facts that arise in maximum likelihood estimation of parameters of a Gaussian distribution.

Recall that a symmetric matrix $M \in \mathbb{R}^{n\times n}$ is positive semidefinite if, for all $x \in \mathbb{R}^n$, $x'Mx \geq 0$. It is positive definite if $x'Mx = 0$ implies $x = 0$. Notice that any covariance matrix is symmetric and positive semidefinite. Indeed, if Σ is the covariance of the random vector X, then $a'\Sigma a$ is the variance of $a'X$.

Lemma 1. *Let $M \in \mathbb{R}^{p\times p}$ be a positive definite symmetric matrix, $W \in \mathbb{R}^{d\times d}$ a positive semidefinite symmetric matrix, and $N \in \mathbb{R}^{p\times d}$. Then*

$$\arg \min_{A\in\mathbb{R}^{p\times d}} \operatorname{tr}\left(W\left(A'MA - N'A - A'N\right)\right) = M^{-1}N.$$

As an example, consider least squares linear regression, with data $Y \in \mathbb{R}^{n\times d}$, $X \in \mathbb{R}^{n\times p}$ and linear parameters $\Theta \in \mathbb{R}^{p\times d}$. Then the sum of the squared errors is

$$\operatorname{tr}\left((Y - X\Theta)'(Y - X\Theta)\right) = \operatorname{tr}\left(Y'Y + \Theta'X'X\Theta - Y'X\Theta - \Theta'X'Y\right),$$

and so the minimizing Θ is $(X'X)^{-1}(X'Y)$.

Lemma 2. *If $P, S \in \mathbb{R}^{n\times n}$ are symmetric positive definite matrices,*

$$\ln|P| + \operatorname{tr}\left(P^{-1}S\right) \geq \ln|S| + \operatorname{tr}\left(S^{-1}S\right).$$

Thus, the left hand side of the inequality is minimized by setting $P = S$.

The two lemmas can be proved using the matrix derivative identities

$$\frac{\partial}{\partial A}\operatorname{tr}(AB) = B',$$
$$\frac{\partial}{\partial A}\log|A| = A^{-T};$$

see Chapter 13 in the text. We give alternative proofs, based on the spectral representation of positive semidefinite symmetric matrices. For the first lemma, the proof directly exploits the intuition that the criterion we are minimizing is quadratic in the parameters, and so for an appropriate choice of basis the optimization problem decouples and becomes trivial. For the second lemma, we see that the criterion depends only on the eigenvalues of a certain positive semidefinite matrix, so the optimization problem again decouples.

Spectral representation of psd symmetric matrices

For a positive semidefinite symmetric matrix $A \in \mathbb{R}^{n\times n}$, we can write

$$A = \sum_{i=1}^{n} \lambda_i u_i u_i' = U\Lambda U',$$

where the $\lambda_i \geq 0$ are the eigenvalues of A, the orthonormal vectors $u_i \in \mathbb{R}^d$ are the eigenvectors, $U \in \mathbb{R}^{n\times n}$ has ith column u_i, and $\Lambda = \text{diag}(\lambda_i) \in \mathbb{R}^{n\times n}$ is the diagonal matrix with ith diagonal entry λ_i.

We can define the square root of any positive semidefinite symmetric matrix. Indeed, since the u_i are orthonormal, we have $U'U = I_n$, so we can write

$$A = U\Lambda U' = \left(U\text{diag}(\sqrt{\lambda}_i)U'\right)\left(U\text{diag}(\sqrt{\lambda}_i)U'\right) =: A^{1/2}A^{1/2}.$$

Proof of Lemma 1

To prove the lemma, we 'complete the square,' expressing the quadratic objective in the form of a positive semidefinite BWB', and then use the spectral representation to conclude that the optimal B is the zero matrix:

$$\begin{aligned}
&\arg\min_{\Theta} \text{tr}\left(W\left(\Theta' M\Theta - N'\Theta - \Theta' N\right)\right) \\
&= \arg\min_{\Theta} \text{tr}\left(W\left(M^{1/2}\Theta - M^{-1/2}N\right)'\left(M^{1/2}\Theta - M^{-1/2}N\right)\right) \\
&= \arg\min_{\Theta} \text{tr}\left(\left(M^{1/2}\Theta - M^{-1/2}N\right)W\left(M^{1/2}\Theta - M^{-1/2}N\right)'\right).
\end{aligned}$$

Notice that, for any B, the matrix BWB' is positive semidefinite and symmetric, because we can write

$$x'BWB'x = \left(W^{1/2}B'x\right)'\left(W^{1/2}B'x\right) \geq 0,$$

and so its trace (which is equal to the sum of its eigenvalues) is non-negative. If we can choose $B = 0$, the trace is clearly minimized. Thus, the optimal value of Θ is $M^{-1}N$.

Proof of Lemma 2

Using the identities $|AB| = |A||B|$ and $|A^{-1}| = 1/|A|$, we can write

$$\ln|P| + \text{tr}\left(P^{-1}S\right) = -\ln|P^{-1}S| + \text{tr}(P^{-1}S) + \ln|S|.$$

Let λ_i be the (nonnegative) eigenvalues of the symmetric psd matrix $P^{-1}S$. Since the determinant is the product of eigenvalues and the trace the sum of eigenvalues, we have

$$\begin{aligned}
-\ln|P^{-1}S| + \text{tr}(P^{-1}S) &= -\ln\left(\prod_i \lambda_i\right) + \sum_i \lambda_i \\
&= \sum_i \left(\lambda_i - \ln\lambda_i\right).
\end{aligned}$$

It is easy to see that $J(\lambda_i) = \lambda_i - \ln\lambda_i$ is convex and has a minimum at $\lambda_i = 1$. Thus, the criterion is minimized when all eigenvalues of $P^{-1}S$ are equal to 1, that is, when $P^{-1}S$ is the identity matrix.