The Akaike Information Criterion for Model Order Selection in Gaussian Data

Navid Shokouhi

The Akaike information criterion (AIC) is the most popular objective function for model selection. AIC uses the principle of Maximum Likelihood, which is equivalent to minimizing the Kullback Leibler divergence (KL-divergence) between the true probability density function (pdf), $f(\mathbf{x}_1, \ldots, \mathbf{x}_N)$, and the parametric pdf corresponding to the model, $\mathcal{M}$, $f(\mathbf{x}_1, \ldots, \mathbf{x}_N|\mathcal{M})$ [1]. This document presents a step-by-step derivation of AIC for N multivariate iid samples of dimension p, $\mathbf{x}_i \in \mathbb{R}^p$, with Gaussian distribution, $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$. The fact that $\mathbf{x}_i$ are zero-mean does not effect the generality of the problem and is only for convenience. When possible, we will use the $p \times N$ matrix $\mathbf{X} = [\mathbf{x}_1 \ldots \mathbf{x}_N]$ to represent the data.

The conditions imposed on the data are the same as those described in [2]. It is assumed that $\mathbf{\Sigma}$ has q large eigenvalues ($\lambda_1 > \lambda_2 > \cdots > \lambda_q$). The remaining $p - q$ eigenvalues are assumed to be equal $\lambda_q > \lambda_{q+1} = \lambda_{q+2} = \ldots \lambda_p$. The value q is called the model order. AIC is described for a much broader class of data and is defined as:

$$AIC(k) = -2\mathcal{L}(\mathbf{X}|\mathcal{M}_k) + 2\gamma_k \tag{1}$$

where $\mathcal{L}(\mathbf{X}|\mathcal{M}_k)$ denotes the data log-likelihood for the model $\mathcal{M}_k$, which corresponds to k large eigenvalues. The variable γ_k is the degree of freedom of $\mathcal{M}_k$. Under the right conditions, $AIC(k)$ has a minimum at $k = q$. For iid samples, $\mathcal{L}(\mathbf{X})$, can be separated as:

$$\mathcal{L}(\mathbf{X}|\mathcal{M}_k) = \log(f(\mathbf{x}_1|\mathcal{M}_k)) + \cdots + \log(f(\mathbf{x}_N|\mathcal{M}_k)) = \sum_{i=1}^{N} \log(f(\mathbf{x}_i|\mathcal{M}_k)) \tag{2}$$

for Gaussian distributions, using the trace operation, $tr(.)$:

$$f(\mathbf{x}_i|\mathcal{M}_k) = \frac{1}{|2\pi\mathbf{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2} tr\left(\mathbf{\Sigma}^{-1}\mathbf{x}_i\mathbf{x}_i^T\right)\right) \tag{3}$$

$$\implies \mathcal{L}(\mathbf{X}|\mathcal{M}_k) = -\frac{pN}{2}\log(2\pi) - \frac{N}{2}\log\left(|\mathbf{\Sigma}|\right) - \frac{N}{2} tr(\mathbf{\Sigma}^{-1}\mathbf{S}), \tag{4}$$

where $\mathbf{S} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i\mathbf{x}_i^T$ is the sample covariance matrix. In order to calculate $\mathcal{L}(\mathbf{X}|\mathcal{M}_k)$, we must know the covariance, $\mathbf{\Sigma}$. Since $\mathbf{\Sigma}$ is generally not available, we can use the maximum likelihood estimate of $\mathbf{\Sigma}$, which is represented through its eigenvalue decomposition, $\mathbf{\Sigma} = \mathbf{U}\mathbf{D}\mathbf{U}^T$.

$$\mathcal{L}(\mathbf{X}|\mathcal{M}_k) = -\frac{Np}{2}\log(2\pi) - \frac{N}{2}\log\left|\mathbf{U}\begin{pmatrix}\mathbf{\Lambda} & \mathbf{0} \\ \mathbf{0} & \sigma^2\mathbf{I}_{p-k}\end{pmatrix}\mathbf{U}^T\right| - \frac{N}{2} tr\left(\mathbf{S}\mathbf{U}\begin{pmatrix}\mathbf{\Lambda}^{-1} & \mathbf{0} \\ \mathbf{0} & \sigma^{-2}\mathbf{I}_{p-k}\end{pmatrix}\mathbf{U}^T\right) \tag{5}$$

where $\mathbf{\Lambda}$ is the $k \times k$ diagonal matrix containing the first k eigenvalues and σ^2 is the value assigned to the last $p-k$ eigenvalues. Using the fact that $\mathbf{U}$ is unitary (i.e., $|\mathbf{U}| = |\mathbf{U}^T| = 1$):

$$\mathcal{L}(\mathbf{X}|\mathcal{M}_k) = -\frac{Np}{2}\log(2\pi) - \frac{N}{2}\log\left(\prod_{j=1}^{k}\lambda_j \prod_{j=k+1}^{p}\sigma^2\right) - \frac{N}{2} tr\left(\mathbf{S}\mathbf{U}\begin{pmatrix}\mathbf{\Lambda}^{-1} & \mathbf{0} \\ \mathbf{0} & \sigma^{-2}\mathbf{I}_{p-k}\end{pmatrix}\mathbf{U}^T\right) \tag{6}$$

$$= -\frac{Np}{2}\log(2\pi) - \frac{N}{2}\sum_{j=1}^{k}\log(\lambda_j) - \frac{N}{2}(p-k)\log(\sigma^2) - \frac{N}{2} tr\left(\mathbf{S}\mathbf{U}\begin{pmatrix}\mathbf{\Lambda}^{-1} & \mathbf{0} \\ \mathbf{0} & \sigma^{-2}\mathbf{I}_{p-k}\end{pmatrix}\mathbf{U}^T\right) \tag{7}$$

Maximizing $\mathcal{L}(\mathbf{X}|\mathcal{M}_k)$ with respect to λ_j, σ^2, and $\mathbf{U}$, will result in the maximum likelihood estimation. Here, we will only maximize with respect to λ_j and σ^2. Maximizing with respect to the eigenvectors (i.e., columns of $\mathbf{U}$) is a fairly tedious exercise. Without derivation, we use [**?**], which shows that the maximum likelihood of $\mathbf{U}$ is $\mathbf{V}$, where $\mathbf{V}$ is the matrix of eigenvectors of the sample covariance matrix, $\mathbf{S}$).

for λ_j:

$$\frac{\partial\mathcal{L}(\mathbf{X}|\mathcal{M}_k)}{\partial\lambda_j} = 0 - \frac{N}{2}\frac{1}{\lambda_j} - 0 - \frac{N}{2} tr\left(\mathbf{S}\left[\frac{-1}{\lambda_j^2}\mathbf{u}_j\mathbf{u}_j^T\right]\right) = 0, \tag{8}$$

where $\mathbf{u}_j$ is the j^{th} column of $\mathbf{U}$.

$$-\frac{1}{\lambda_j} + \frac{1}{\lambda_j^2}\mathbf{u}_j^T\mathbf{S}\mathbf{u}_j = 0, \tag{9}$$

which results in $\lambda_j \mathbf{u}_j = \mathbf{S}\mathbf{u}_j$. Therefore, the maximum likelihood estimate of λ_j, by definition, is the j^{th} eigenvalue of $\mathbf{S}$ ($\hat{\lambda_j} = l_j$).

for σ^2:

$$\frac{\partial \mathcal{L}(\mathbf{X}|\mathcal{M}_k)}{\partial \sigma^2} = 0 - 0 - \frac{N}{2}\frac{1}{\sigma^2} - \frac{N}{2} tr\left(\mathbf{S}\left[\frac{-1}{(\sigma^2)^2}\sum_{j=k+1}^{p}\mathbf{u}_j\mathbf{u}_j^T\right]\right) = 0$$

$$\implies (p-k)\frac{1}{\sigma^2} = \frac{1}{(\sigma^2)^2} tr(\mathbf{S}\sum_{j=k+1}^{p}\mathbf{u}_j\mathbf{u}_j^T)$$

$$\implies \sigma^2 = \frac{1}{p-k}\sum_{j=k+1}^{p} tr(\mathbf{S}\mathbf{u}_j\mathbf{u}_j^T)$$

$$\implies \sigma^2 = \frac{1}{p-k}\sum_{j=k+1}^{p} \mathbf{u}_j^T\mathbf{S}\mathbf{u}_j \tag{10}$$

where $\mathbf{u}_j^T\mathbf{S}\mathbf{u}_j$ gives the last $p-k$ eigenvalues of $\mathbf{S}$. Therefore, the maximum likelihood estimate of σ^2 is:

$$\hat{\sigma^2} = \frac{1}{p-m}\sum_{j=k+1}^{p} l_j \tag{11}$$

The final step of deriving $AIC(k)$ is to insert the maximum likelihood estimates in $\mathcal{L}(\mathbf{X}|\mathcal{M}_k)$ in Eq. (6).

$$\mathcal{L}(\mathbf{X}|\mathcal{M}_k) = -\tfrac{Np}{2}\log(2\pi) - \tfrac{N}{2}\log\left(\prod_{j=1}^{k}\lambda_j\prod_{j=k+1}^{p}\hat{\sigma}^2\right) - \tfrac{N}{2} tr\left(\mathbf{V}\mathbf{L}\mathbf{V}^T\mathbf{V}\hat{\mathbf{D}}^{-1}\mathbf{V}^T\right)$$

$$= -\tfrac{Np}{2}\log(2\pi) - \tfrac{N}{2}\log\left(\prod_{j=1}^{k} l_j \prod_{j=k+1}^{p}\left(\tfrac{1}{p-k}\sum_{j=k+1}^{p} l_j\right)\right) - \tfrac{N}{2} tr\left(\mathbf{V}\begin{pmatrix}\boldsymbol{I}_k & \mathbf{0} \\ \mathbf{0} & \begin{pmatrix}\hat{\sigma^{-2}}l_1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \hat{\sigma^{-2}}l_2\end{pmatrix}\end{pmatrix}\mathbf{V}^T\right) \tag{12}$$

$$= -\tfrac{Np}{2}\log(2\pi) - \tfrac{N}{2}\log\left(\prod_{j=1}^{k} l_j \prod_{j=k+1}^{p}\left(\tfrac{1}{p-k}\sum_{j=k+1}^{p} l_j\right)\right) - \tfrac{N}{2}(k + \tfrac{1}{\hat{\sigma^2}}\sum_{j=k+1}^{p} l_j) \tag{13}$$

where we have used the property of the trace operation, $tr(\mathbf{V}\mathbf{S}\mathbf{V}^T) = tr(\mathbf{S}\mathbf{V}^T\mathbf{V})$. Finally, from Eq. (11), we have $\frac{1}{\hat{\sigma^2}}\sum_{j=k+1}^{p} l_j = p-k$. Therefore,

$$\mathcal{L}(\mathbf{X}|\mathcal{M}_k) = -\frac{Np}{2}\left(\log(2\pi)+1\right) - \frac{N}{2}\log\left(\left(\frac{1}{p-k}\sum_{j=k+1}^{p} l_j\right)^{(p-k)}\prod_{j=1}^{k} l_j\right) \tag{14}$$

Compare this with the result presented in [2].

REFERENCES

[1] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *Selected Papers of Hirotugu Akaike*. Springer, 1998, pp. 199–213.

[2] M. Wax and T. Kailath, "Detection of signals by information theoretic criteria," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 387–392, 1985.