



Why

What if we use the principle of maximum likelihood to select the maximum likelihood normal density which factors according to a tree?

Definition

A *maximum likelihood tree normal* of a dataset in \mathbf{R}^d is a multivariate normal density that factors according to a tree and maximizes the likelihood of the dataset.

Results

Proposition 1. *Let $D = (x^1, \dots, x^n)$ be a dataset in \mathbf{R}^d . A normal density is a maximum likelihood tree normal of D if and only if it is an optimal tree approximator of the empirical normal density of D .*

Proof. First, let $f : \mathbf{R}^d \rightarrow \mathbf{R}$ be a normal density.

First, express the log likelihood of f on a record x^k by

$$\log f(x^k) = -\frac{1}{2}(x^k - \mu)^\top \Sigma^{-1}(x^k - \mu) - \frac{1}{2} \log \det \Sigma - \frac{d}{2} \log 2\pi.$$

Second, use the trace to rewrite the quadratic form

$$-\frac{1}{2} \operatorname{tr}(\Sigma^{-1}(x^k - \mu)(x^k - \mu)^\top).$$

Third, use these two, and the linearity of trace to express the average negative log likelihood by

$$-\frac{1}{n} \sum_{k=1}^n \log f(x^k) = \frac{1}{2} \operatorname{tr}(\Sigma^{-1}(\frac{1}{n} \sum_{i=1}^n (x^k - \mu)(x^k - \mu)^\top)) + \frac{1}{2} \log \det \Sigma + \frac{d}{2} \log 2\pi.$$

Fourth, use matrix calculus (or the derivation in Proposition 1 of Multivariate Normal Maximum Likelihood) to see that, for a minimizer of the negative average log likelihood, the mean must be $\frac{1}{n} \sum_{i=1}^n x^k$.

Fifth, recognize the empirical covariance matrix $\frac{1}{n} \sum_{k=1}^n (x^k - \mu)(x^k - \mu)^\top$; denote it by S .

Sixth, change variables with $P = \Sigma^{-1}$ and express

$$\log(\det(\Sigma)) = \log(\det(P^{-1})) = \log((\det P)^{-1}) = -\log(\det(P))$$

Seventh, write the likelihood in simplified form (using circulant property of trace)

$$\frac{1}{2} \text{tr}(SP) - \frac{1}{2} \log \det P - \frac{d}{2} \log 2\pi$$

Eighth, drop the constant and prefactors:

$$\text{tr}(SP) - \log \det P$$

Ninth, if g is a normal with then the tree density approximation objective is the same equivalent to

$$d(g, f) = h(g, f) - h(f) \sim h(g, f) = - \int_{\mathbf{R}^d} g \log f.$$

TODO: Extra, the let g be normal and f be normal. The tree normal approximation problem

$$d(g, f) \sim h(g, f) = - \int_{\mathbf{R}^d} g \log f.$$

The log of f is

$$-\frac{1}{2} \text{tr}(\Sigma_f \int_{\mathbf{R}^d} (x - \mu_f)(x - \mu_f)^\top dx) - \frac{1}{2} \log \det \Sigma_f^{-1} - \frac{d}{2} \log 2\pi$$

Since the set of optimal solutions for both optimization is contained in the set of normal densities which match on the mean we can assume that $\mu_f = \mu_g$. So the approximation objective is equivalent to

$$\text{tr} P_f \Sigma_g - \log \det P_f,$$

which is exactly the maximum likelihood objective.

Thus, a solution is a maximum likelihood tree normal of a dataset if and only if it is an optimal tree approximator of the empirical normal density of the dataset.

□

