NN

Frances Lin

Dec 2022

NN (Nearest Neighbor)

from Datta et al. (2016)

Let $\mathcal{S} = \{s_1, s_2, ..., s_k\}$ be a fixed collection of distinct locations in $\mathcal{D} \subseteq \mathcal{R}^d$ ($\mathcal{S}$ is referred to as the *reference set*), then by the chain rule, joint density of $w_{\mathcal{S}}$ can be expressed as a product of conditional densities. That is,

$$p(w_{\mathcal{S}}) = p(w(s_1))p(w(s_2)|w(s_1)) \cdots p(w(s_k)|w(s_{k-1}), ..., w(s_1))$$
$$= \prod_{i=1}^{k} p(w(s_i)| \cap_{j=1}^{k-1} w(s_j)).$$

Next replace the right-hand side of... with smaller, carefully chosen, conditioning sets of size at most m, where $m \ll k$ (see, e.g., ...), then, for every $s_i \in \mathcal{S}$, a smaller conditioning sets $N(s_i) \subset \mathcal{S} \setminus \{s_i\}$ is used to construct a new density

$$\tilde{p}(w_{\mathcal{S}}) = \prod_{i=1}^{k} p(w(s_i)|w_{N(s_i)}),$$

where $w_{N(s_i)}$ is the vector of $w(s)$ over $N(s_i)$.

The pair $\{\mathcal{S}, N_{\mathcal{S}}\}$ can be viewed as a directed graph $\mathcal{G}$, where $\mathcal{S}(= \{s_1, s_2, ..., s_k\})$ is the set of nodes and $N_{\mathcal{S}}(= \{N(s_i); i = 1, 2, ..., k\})$ is the set of directed edges. $N(s_i)$ denotes the set of directed neighbors of s_i ($N(s_i)$ is referred to as the *neighbor set* for s_i). If $\mathcal{G}$ is a directed acyclic graph, then $\tilde{p}(w_{\mathcal{S}})$ is a proper multivariate joint density (see Appendix A1 of Datta et al., 2016). In addition, for a very general class of neighboring sets, $\tilde{p}(w_{\mathcal{S}})$ is a joint density of a multivariate Gaussian distribution with a sparse precision matrix $\tilde{C}_{\mathcal{S}}^{-1}$. More specifically, let $C_{N(s_i)}$ be the covariance matrix of $w_{N(s_i)}$ and $C_{s_i,N(s_i)}$ be the cross-covariance matrix between $w(s_i)$ and $w_{N(s_i)}$, then $\tilde{p}(w_{\mathcal{S}})$ is a multivariate Gaussian density with covariance matrix $\tilde{C}_{\mathcal{S}} = {B_{\mathcal{S}}}^{-1} F_{\mathcal{S}}^{-1} B_{\mathcal{S}}$ and

$$\tilde{p}(w_{\mathcal{S}}) = \prod_{i=1}^{k} N(w(s_i)|B_{s_i} w_{N(s_i)}, F_{s_i}),$$

where $B_{s_i} = C_{s_i,N(s_i)} C_{N(s_i)}^{-1}$ and $F_{s_i} = C(s_i, s_i) - C_{s_i,N(s_i)} C_{N(s_i)}^{-1} C_{N(s_i),s_i}$. This is because, by the theorem, if $p(w_{\mathcal{S}}) = N(w_{\mathcal{S}}|0, C_{\mathcal{S}})$, then $w(s_i)|w_{N(s_i)} \sim N(B_{s_i} w_{N(s_i)}, F_{s_i})$ (see Appendix A2 of Datta et al., 2016).

..... $\tilde{p}(w_{\mathcal{S}})$ is referred to as the nearest neighbor density of $w_{\mathcal{S}}$.

from Finley et al. (2019)

That is, the underlying idea of the NNGP models is similar to that of the graphical models. The joint distribution for a random vector w can be viewed as a directed acyclic graph (DAG). More specifically, $p(w) = p(w_1, w_2, ...w_n)$ can be written as

$$p(w) = p(w_1) \prod_{i=2}^{n} p(w_i|Pa[i]), \tag{4-1}$$

where $w_i \equiv w(s_i)$ and $Pa[i] = \{w_1, w_2, ...w_{i-1}\}$ is a set of parents of w_i,

or, more explicitly, as

$$p(w) = p(w_1)p(w_2|w_1)p(w_3|w_1, w_2) \cdots p(w_n|w_1, w_2, ...w_{i-1}) \tag{4-2}$$

(Datta et al., 2017). Sparse models for w can be constructed by shrinking the size of $Pa[i]$. .....

..... The multivariate Gaussian density $N(w|0, C)$ (or $w \sim N(0, C(\theta))$) in (2) can be written as a linear model

$$w_1 = 0 + \eta_1,$$

$$w_2 = a_{21}w_1 + \eta_2,$$

$$w_i = a_{i1}w_1 + a_{i2}w_2 + \cdots + a_{i,i-1}w_{i-1} + \eta_i, \text{ for } i = 2, ..., n, \tag{4-3}$$

or, more explicitly, as

$$\begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ \vdots \\ w_n \end{bmatrix} = \begin{bmatrix} 0 & 0 & \cdots & 0 & 0 \\ a_{21} & 0 & \cdots & 0 & 0 \\ a_{31} & a_{32} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn_1} & 0 \end{bmatrix} + \begin{bmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \\ \vdots \\ \eta_n \end{bmatrix} \tag{4-4}$$

(Datta et al., 2017),

or, more compactly, as

$$w = Aw + \eta, \tag{4-5}$$

where A is nxn strictly lower-triangular and $\eta \sim N(0, D)$ with $D = (d_1, d_2, ..., d_n)$ is diagonal. It follows that $I - A$ is nonsingular and, by the Cholesky factorization (Cholesky decomposition), a covariance matrix C can be factorized into a product $C = (I - A)^{-1}D(I - A)^{-T}$, where for any matrix M, M^{-T} refers to the inverse of its transpose.

However, the Cholesky factorization for the full GP covariance C does not offer any computational advantages. Instead, the sparsity was introduced through graphical models (Datta et al., 2017).

To construct a sparse precision matrix, start with a dense nxn covariance matrix C and construct a sparse strictly lower-triangular matrix A with no more than $m(\ll n)$ nonzero entries in each row and the diagonal matrix D, then the matrix $\tilde{C} = (I - A)^{-1}D(I - A)^{-T}$ is a covariance matrix and its inverse $\tilde{C}^{-1} = (I - A)^T D^{-1}(I - A)$ is sparse. This leads to the latent NNGP model in the section below.

NNGP can also be viewed as as a special case of a Gaussian Markov Random Field (GMRF; Rue and Held 2005, as cited in Finley et al., 2021).