1 Abstract & Introduction

The authors introduce a *multiscale model*, i.e. a model that decomposes the overall structure of an object under study according to its component structures at different scales of spatial and/or temporal resolution. The key assumption of these models is that there exists a set of hierarchically defined partitions, corresponding to successive aggregations of the space. In this paper, they study the effect of scale by studying the data likelihood in each part of the hierarchical partitions.

2 Aggregation & The Likelihood

2.1 The Spatial Model

We wish to take measurements on the domain $D \subset \mathbb{R}^n$ of an underlying mean process $\{\mu(s) : s \in D\}$. We assume that there is some granularity or aggregation beyond which the data and μ cannot be resolved. The aggregation at this granularity is defined by the parititon $\{B_1, \ldots, B_n\}$ of D, where $B_k \subset D, B_k \cap B_{k'} = \emptyset$ for $k \neq k'$, and $\cup_k B_k = D$. We denote the measurement as Y_k and this is linked to the mean process μ by $E[Y_k] = \mu_k \equiv \mu(B_k) = \int_{B_k} \mu(s)ds$.

We can defined coarser aggregations than $\{B_1, \ldots, B_n\}$ as a set $\mathscr{B}$ of $J+1$ nested partitions of D, for integer $J \geq 1$. The original, and finest, partition is now denoted $\{B_{J,1}, \ldots, B_{J,n_J}\}$. For each scale $j = 0, \ldots, J-1$ a collection of the elements from the jth partition $\{B_{j,k}\}_{k=1}^{n_j}$ is defined such that:

- the collection is a proper partition of D
- It must be possible to express every $B_{j,k}$ as the union of a unique set of elements in the partition at scale $j+1$, $B_{j,k} = \cup_{k' \in ch(k)} B_{j+1,k'}$, where $ch(k)$ is the collection of spatial indices of the children of $B_{j,k}$.

Now consider the kth element of the jth partition $B_{j,k}$. The measurement corresponding to this element $Y_{j,k}$ is the sum of the measurements corresponding to the children of $B_{j,k}$

$$Y_{j,k} = \sum_{k' \in ch(k)} Y_{j+1,k'} \quad \text{and} \quad \mu_{j,k} = \sum_{k' \in ch(k)} \mu_{j+1,k'} \tag{1}$$

Information is lost in the process of aggregating from finer partitions to coarser partitions. We would like to be able to quantify how much information. We will use the likelihood as a measure of the information.

2.2 Multiscale Likelihood Factorizations

We can define the relationship between a collection of the measurements on the finest partition that (uniquely) form all the children of an element of the next finest partition with that element as follows. Let $n_{ch(k)}$ denote the total number of children minus one, then

$$P(Y_{J,1}, \ldots, Y_{J,n_{ck(k)}} | \mu_{J,1}, \ldots \mu_{J,n_{ck(k)}}) = P(Y_{J,1}, \ldots, Y_{J,n_{ck(k)}} | Y_{J-1,k}, \mu_{J,1}, \ldots \mu_{J,n_{ck(k)}}) \\ P(Y_{J-1,k} | \mu_{J-1,k})$$

or more concisely

$$P(\mathbf{Y}_{J,ch(k)}|\boldsymbol{\mu}_{J,ch(k)}) = P(\mathbf{Y}_{J,ch(k)}|Y_{J-1,k}, \boldsymbol{\mu}_{J,ch(k)})P(Y_{J-1,k}|\mu_{J-1,k}) \tag{2}$$

or even more concisely

$$P_{\boldsymbol{\mu}}(\mathbf{Y}_{J,ch(k)}) = P_{\boldsymbol{\mu}}(\mathbf{Y}_{J,ch(k)}|Y_{J-1,k})P_{\boldsymbol{\mu}}(Y_{J-1,k}) \tag{3}$$

Remembering that the $Y_{J,k}$ are conditionally independent given $\mu_{J,k}$ we can write

$$\prod_{k'=1}^{n_{ch(k)}} P_{\mu}(Y_{J,k'}) = P_{\mu}(Y_{J-1,k}) \prod_{k'=1}^{n_{ch(k)}} P_{\mu}(Y_{J,k'}|Y_{J-1,k}) \tag{4}$$

and we can extend that to each of the parent child groups

$$\prod_{k=1}^{n_J} P_{\mu}(Y_{J,k}) = \prod_{k=1}^{n_{J-1}} P_{\mu}(Y_{J-1,k}) \prod_{k=1}^{n_{J-1}} P_{\boldsymbol{\mu}}(\mathbf{Y}_{J,ch(k)}|Y_{J-1,k}) \tag{5}$$

Note that $\prod_{k=1}^{n_J} P_{\mu}(Y_{J,k})$ is the likelihood of the measurements/data at scale J. Further note that we have the likelihood of the data at scale J on the left hand side of equation 5, and the likelihood of the data at scale $J-1$ on the right hand side. So the second factor on the right hand side of equation 5 must be information lost due to aggregation.

We can factorise the likelihood at scale $J-1$, $J-2$, etc., in the same fashion which gives the fully factorized likelihood:

$$\prod_{k=1}^{n_J} P_{\mu}(Y_{J,k}) = \left[\prod_{k=1}^{n_0} P_{\mu}(Y_{0,k})\right] \left[\prod_{j=0}^{J-1} \prod_{k=1}^{n_j} P_{\boldsymbol{\mu}}(\mathbf{Y}_{j+1,ch(k)}|Y_{j,k})\right] \tag{6}$$

In graph theory, a graph that has the *directed local Markov* property must have a factorisation like that described in equation 6.

2.3 Likelihood Factorisation for Continuous and Count Data

2.3.1 Gaussian case

If the measurements on the finest partition J are sampled from a Gaussian distribution $\mathbf{Y}_J|\boldsymbol{\mu}_J \sim \mathcal{N}(\boldsymbol{\mu}_J, \Sigma_J)$ it can be shown that:

$$P(\mathbf{Y}_J|\boldsymbol{\mu}_J, \Sigma_J) = \mathcal{N}(\mathbf{Y}_0|\boldsymbol{\mu}_0, \Sigma_0) \prod_{j=0}^{J-1} \prod_{k=1}^{n_J} P(\mathbf{Y}_{j+1,ch(k)}|Y_{j,k}, \boldsymbol{\omega}_{j,k}, \Omega_{j,k}) \tag{7}$$

where $\mathbf{Y}_0 \sim \mathcal{N}(\boldsymbol{\mu}_0, \Sigma_0)$ and

$$P(\mathbf{Y}_{j+1,ck(k)}|Y_{j,k}, \boldsymbol{\omega}_{j,k}, \Omega_{j,k}) = \mathcal{N}(\boldsymbol{\nu}_{j,k}Y_{j,k} + \boldsymbol{\omega}_{j,k}, \Omega_{j,k}) \tag{8}$$

with $\boldsymbol{\nu}_{j,k} = \frac{\boldsymbol{\sigma}^2_{j+1,ch(k)}}{\sigma^2_{j,k}}$ and

$$\boldsymbol{\omega}_{j,k} = \boldsymbol{\mu}_{j+1,ch(k)} - \boldsymbol{\nu}_{j,k}\mu_{j,k} \tag{9}$$

and

$$\Omega_{j,k} = \Sigma_{j+1,ch(k)} - \frac{1}{\sigma^2_{j,k}} \boldsymbol{\sigma}^2_{j+1,ch(k)} \left[\boldsymbol{\sigma}^2_{j+1,ch(k)}\right]^T \tag{10}$$

Note that the multiscale means $(\boldsymbol{\mu}_0, \boldsymbol{\omega}_0, \ldots, \boldsymbol{\omega}_{J-1})$ are a reparametrisation of the original means $\boldsymbol{\mu}_J$. So knowledge of the multiscale parameters is equivalent to knowledge of the original mean variables, but only knowledge of the multiscale parameters allows us to see the effects of scale on the analysis.

2.3.2 Poisson case

If the measurements on the finest partition J are sampled from a Poisson distribution $\mathbf{Y}_J|\boldsymbol{\mu}_J \sim Poiss(\boldsymbol{\mu}_J)$, then it can be shown that

$$P(\mathbf{Y}_J|\boldsymbol{\mu}_J) = Poiss(\boldsymbol{\mu}_0) \prod_{j=0}^{J-1} \prod_{k=1}^{n_j} P(\mathbf{Y}_{j+1,ch(k)}|Y_{j,k}, \boldsymbol{\omega}_{j,k}) \tag{11}$$

where $\mathbf{Y}_{j+1,ch(k)}|Y_{j,k}, \boldsymbol{\omega}_{j,k} \sim Mult(Y_{j,k}; \boldsymbol{\omega}_{j,k})$, the multinomial distribution with parameter components

$$\boldsymbol{\omega}_{j,k} = \frac{\boldsymbol{\mu}_{j+1,ch(k)}}{\mu_{j,k}} \tag{12}$$

Note that each element of $\boldsymbol{\omega}_{j,k}$ is between 1 and 0, and the some of the elements will equal 1. So the requirements for the multinomial distribution are satisfied.

3 Estimation

Here we're estimating the $\boldsymbol{\mu}_J$ given the $\mathbf{Y}_J$. We take a Bayesian approach. First specifying priors $P(\boldsymbol{\theta})$, then optimising the posterior distribution $P(\boldsymbol{\theta}|\mathbf{X}) \propto P(\mathbf{X}|\boldsymbol{\theta})P(\boldsymbol{\theta})$.

3.1 Prior Distributions

In this section the authors consider the case where the original measurements $Y_{J,k}$ are proportional to the area of the original partition elements. That is, the area of $B_{J,k}$ is $A_{J,k}$, and $c_1 A_{J,k} = \mu_{J,k}$, and $c_2 A_{J,k} = \sigma^2_{J,k}$, where $c_1 > 0$, and $c_2 > 0$ are just constants.

The consequence of this assumption in the Guassian case is,

$$\boldsymbol{\nu}_{j,k} = \frac{\boldsymbol{\sigma}_{j+1,ch(k)}}{\sigma^2_{j,k}} = \frac{\mathbf{A}_{j+1,ch(k)}}{A_{j,k}}$$

and

$$\begin{aligned}
\boldsymbol{\omega}_{j,k} &= \boldsymbol{\mu}_{j+1,ch(k)} - \boldsymbol{\nu}_{j,k}\mu_{j,k} \\
&= c_1 \mathbf{A}_{j+1,ch(k)} - \frac{\mathbf{A}_{j+1,ch(k)}}{A_{j,k}} c_1 A_{j,k} \\
&= 0
\end{aligned}$$

Therefore the prior

$$P(\boldsymbol{\omega}_{j,k}|\boldsymbol{\Phi}_{j,k}) = \mathcal{N}(\mathbf{0}, \boldsymbol{\Phi}_{j,k}) \tag{13}$$

is chosen, where $\mathbf{\Phi}_{j,k}$ is an $n_{ch(k)} \times n_{ch(k)}$ covariance matrix. The Gaussian form of this prior is convenient because the conjugate of a Gaussian distribution is another Gaussian distribution.

For the Poisson case we have

$$\begin{aligned}\boldsymbol{\omega}_{j,k} = \frac{\boldsymbol{\mu}_{j+1,ch(k)}}{\mu_{j,k}} &= \frac{c_1}{c_1 A_{j,k}} \begin{pmatrix} A_{j+1,1} \\ \vdots \\ A_{j+1,n_{ch(k)}+1} \end{pmatrix} \\ &= \frac{1}{A_{j,k}} \begin{pmatrix} A_{j+1,1} \\ \vdots \\ A_{j+1,n_{ch(k)}+1} \end{pmatrix}\end{aligned}$$

Remembering that $n_{ch(k)}$ is defined as the number of children of $B_{j,k}$ *minus one*, consider one element of this vector,

$$\frac{A_{j+1,1}}{A_{j,k}} = \frac{A_{j+1,1}}{A_{j+1,1} + \cdots + A_{j+1,n_{ch(k)}+1}} = \frac{1}{1 + \frac{A_{j+1,2}}{A_{j+1,1}} + \cdots + \frac{A_{j+1,n_{ch(k)}+1}}{A_{j+1,1}}}$$

If we assume that all the child regions have the same area, then

$$\begin{aligned}\frac{A_{j+1,1}}{A_{j,k}} &= \frac{1}{1 + n_{ch(k)}} \\ \implies \boldsymbol{\omega}_{j,k} &= \frac{1}{1 + n_{ch(k)}} \mathbf{1}\end{aligned}$$

where $\mathbf{1}$ is a vector of ones of length $n_{ch(k)} + 1$. A sensible choice for the prior in this case is

$$P(\boldsymbol{\omega}_{j,k}) = Dir(\gamma_{j,k}\mathbf{1}) \tag{14}$$

where $\gamma_{j,k} > 0$, and $Dir()$ is the Dirichlet distribution. This means the expected value of the elements of $\boldsymbol{\omega}_{j,k}$ is $\frac{1}{n_{ch(k)}+1}\mathbf{1}$.

4 Further Notes

4.1 Implementation of Gaussian case on a parent with two children

Consider the case of a parent region with mean μ_p and variance σ_p^2. The parent region has two children with means μ_1, μ_2 and σ_1^2, σ_2^2. According to the model if we have measurements for the parent and wish to model measurements for the children, the children are distributed according to equations 8, 9, and 10.

In this case,

$$\begin{aligned}\Omega &= \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix} - \frac{1}{\sigma_p^2} \begin{pmatrix} \sigma_1^2 \\ \sigma_2^2 \end{pmatrix} (\sigma_1^2 \sigma_2^2) \\ &= \begin{pmatrix} \sigma_1^2 \left(1 - \frac{\sigma_1^2}{\sigma_p^2}\right) & -\frac{\sigma_1^2 \sigma_2^2}{\sigma_p^2} \\ -\frac{\sigma_1^2 \sigma_2^2}{\sigma_p^2} & \sigma_2^2 \left(1 - \frac{\sigma_2^2}{\sigma_p^2}\right) \end{pmatrix}\end{aligned}$$

This matrix must be positive definite in order for the distribution in equation 8 to be valid. In order for a 2×2 matrix to be positive definite, the trace must be positive, and the determinant must be positive. As for the determinant,

$$\begin{aligned}
&\det(\Omega) = \sigma_1^2\sigma_2^2\left(\left(1-\frac{\sigma_1^2}{\sigma_p^2}\right)\left(1-\frac{\sigma_2^2}{\sigma_p^2}\right)-\frac{\sigma_1^2\sigma_2^2}{\sigma_p^4}\right)\\
&\det(\Omega) < 0\\
&\implies \sigma_p^2 < \sigma_1^2 + \sigma_2^2
\end{aligned}$$

So when the variance of the parent is less than the sum of the variances of the children, the matrix Ω is not positive definite, and the model of the distribution of the children is invalid. This condition can easily occur. If we sum two Gaussian random variables, the result is another Gaussian random variable, with mean $\mu = \mu_1 + \mu_2$, and variance $\sigma^2 = \sigma_1^2 + \sigma_2^2 + 2\rho\sigma_1\sigma_2$, where ρ is the correlation between the two Gaussian variables. If ρ is negative then the model is invalidated.

4.2 Attempt at deriving Gaussian parameter transformations

Here I attempt to derive the expressions in equations 8, 9, and 10.

In this paper, it is assumed that the $Y_{j,k}$ are conditionally independent given the $\mu_{j,k}$. This assumption gives an expression for the correlation between the parent region and each of the children in terms of the parent and child standard deviations. If the parent region is denoted by p, and the two children are denoted by 1, and 2, then

$$\begin{aligned}
\sigma_{1p} = \sigma_1\sigma_p\rho_{1p} &= E\left[(Y_p-\mu_p)(Y_1-\mu_1)\right]\\
&= E\left[(Y_1+Y_2-\mu_1-\mu_2)(Y_1-\mu_1)\right]\\
&= E\left[(Y_1-\mu_1)^2+(Y_1-\mu_1)(Y_2-\mu_2)\right]\\
&= \sigma_1^2+\rho_{12}\sigma_1\sigma_2
\end{aligned}$$

Because of conditional independence $\rho_{12} = 0$, which gives

$$\begin{aligned}
\sigma_1\sigma_p\rho_{1p} &= \sigma_1^2\\
\rho_{1p} &= \frac{\sigma_1}{\sigma_p}
\end{aligned}$$

This expression holds for child 2 as well, and would hold for all the children in the case where there are more than 2.

Consider the random vector $(Y_p, Y_1, Y_2)^\top$. This vector is sampled from a Gaussian distribution with

$$\begin{aligned}
\boldsymbol{\mu} &= \begin{pmatrix}\mu_p\\ \mu_1\\ \mu_1\end{pmatrix}\\
\Sigma &= \begin{pmatrix}\sigma_1^2 & 0 & \rho_{p1}\sigma_p\sigma_1\\ 0 & \sigma_2^2 & \rho_{p2}\sigma_p\sigma_2\\ \rho_{p1}\sigma_p\sigma_1 & \rho_{p2}\sigma_p\sigma_2 & \sigma_p^2\end{pmatrix} = \begin{pmatrix}\sigma_1^2 & 0 & \sigma_1^2\\ 0 & \sigma_2^2 & \sigma_2^2\\ \sigma_1^2 & \sigma_2^2 & \sigma_p^2\end{pmatrix}
\end{aligned}$$

We want an expression for $P(Y_1, Y_2|Y_p)$. This is equivalent to conditioning some elements of a random Gaussian vector on the other elements. In this situation we get

$$\begin{aligned}\boldsymbol{\mu}_{cond} &= \begin{pmatrix}\mu_1 \\ \mu_2\end{pmatrix} + \frac{1}{\sigma_p^2}\begin{pmatrix}\sigma_1^2 \\ \sigma_2^2\end{pmatrix}(Y_p - \mu_p) \\ &= \boldsymbol{\nu}_p Y_p + \boldsymbol{\omega}_p\end{aligned}$$

where $\boldsymbol{\nu}_p = \frac{1}{\sigma_p^2}\begin{pmatrix}\sigma_1^2 \\ \sigma_2^2\end{pmatrix}$ and $\boldsymbol{\omega}_p = \begin{pmatrix}\mu_1 \\ \mu_2\end{pmatrix} - \boldsymbol{\nu}_p\mu_p$.

As for the covariance matrix, we have

$$\Omega_p = \begin{pmatrix}\sigma_1^2 & 0 \\ 0 & \sigma_2^2\end{pmatrix} - \frac{1}{\sigma_p^2}\begin{pmatrix}\sigma_1^2 \\ \sigma_2^2\end{pmatrix}\begin{pmatrix}\sigma_1^2 & \sigma_2^2\end{pmatrix}$$

5 Variable Directory

For specifying what each varible means, and where it comes from.

$\boldsymbol{\mu}(s)$ 'mean process': Some process operating in each part of our partition. We want to 'obtain knowledge' of this process.

B_k Used to indicate an element of a partition.

Y_k A univariate measurement corresponding to the parition element B_k, linked to the mean process by $E[Y_k] = \boldsymbol{\mu}_k$ where $\boldsymbol{\mu}_k \equiv \boldsymbol{\mu}(B_k) = \int_{B_k} \boldsymbol{\mu}(s)ds$.

$\mathscr{B}$ The set of nested partitions of our space. Each partition that forms a level in our hierarchy will we an element of $\mathscr{B}$.

$\{B_{J,1}, \ldots, B_{J,n_J}\}$ The elements of the finest partition.

$\{B_{j,1}, \ldots, B_{j,n_j}\}$ where $0 \leq j < J$: The elements of a coarser partition. The partition is proper, and each of the elements is the union of a unique collection taken from the next finest parition.

$Pr_{\boldsymbol{\mu}}(\mathbf{Y})$ $Pr_{\boldsymbol{\mu}}(\mathbf{Y}) = Pr(\mathbf{Y}|\boldsymbol{\mu})$. Remember that we assume that the measurements Y_k are conditionally independent given μ_k.

Factorisation across one level In the expression

$$Pr_{\boldsymbol{\mu}}(\mathbf{Y}_{J,ch(k)}) = Pr_{\boldsymbol{\mu}}(\mathbf{Y}_{J,ch(k)}|Y_{J-1,k})Pr_{\boldsymbol{\mu}}(Y_{J-1,k}), \tag{15}$$

$ch(k)$ refers to all but one of the children of $B_{J-1,k}$. If we are given the value of the measurement on the parent, and the value of all but one of the children, then the value of the last child is determined due to our assumptions about the summations of the measurements. Therefore, we only need to define the distribution across all but one of the children. Which child is left out is arbitrary.

$\Sigma_j = diag(\boldsymbol{\sigma_j^2})$ The covariance matrix of a Gaussian distributed measurement across a partition is a diagonal matrix, because of our conditional independence assumption.

$\nu_{j,k}, \omega_{j,k}, \Omega_{j,k}$ Multiscale parameters of the Gaussian distributed partitions.

$$\nu_{j,k} = \frac{\boldsymbol{\sigma}^2_{j+1,ch(k)}}{\sigma^2_{j,k}} \tag{16}$$

$$\omega_{j,k} = \boldsymbol{\mu}_{j+1,ch(k)} - \nu_{j,k}\mu_{j,k} \tag{17}$$

$$\Omega_{j,k} = \Sigma_{j+1,ch(k)} - \frac{1}{\sigma^2_{j,k}}\boldsymbol{\sigma}^2_{j+1,ch(k)}\left[\boldsymbol{\sigma}^2_{j+1,ch(k)}\right]^\top \tag{18}$$

$\omega_{j,k}$ **(Poisson)** Multiscale parameters of the Poisson distributed partitions.

$$\omega_{j,k} = \frac{\boldsymbol{\mu}_{j+1,ch(k)}}{\mu_{j,k}} \tag{19}$$

$A_{j,k}$ The area of the parition element $B_{j,k}$. In a special case of this model, the mean and variance of the data $Y_{j,k}$ is proportional to the area $A_{j,k}$.

$\boldsymbol{\Phi}_{j,k}$ The covariance matrix of the multiscale parameters for the special case of measurements proportional to area.

$\gamma_{j,k}$ The parameter of the Dirichlet distribution used as the prior distribution for the multiscale parameters of the Poisson version of the area proportional model.

$\eta_{j,k}$ A Bernoulli variable used to facilitate mixture models.

$\boldsymbol{\tau}, \boldsymbol{\Phi_0}$ Mean and variance chosen for the prior over $\boldsymbol{\mu}_0$ when attempting to calculate a posterior for $\omega_{j,k}$ in the Gaussian context. The corresponding parameters for the Poisson case are $\boldsymbol{\tau_0^{(1)}}, \boldsymbol{\tau_0^{(2)}}$.