River Stream Flow - Clustering and Modelling

Eleanor Barry, Abigail Verschueren, James Hook

University of Bath

February 2, 2018

Introduction

Data: Daily rainfall and riverflow data (50 years).

Cluster analysis aims to group similar rivers so that inference can be made for rivers without data and for the fitting of random effects.

This method also provides further information about the behaviour of the rivers.

Linear model for rain and flow data

Time series model

$$y_t = X_t w + \epsilon$$

where

- $y_t \in \mathbb{R}$ is the river flow on day t
- $X_t \in \mathbb{R}^m$ is a vector of time series observations e.g $X_t = (r_t, r_{t-1}, \ldots, r_{t-9}, y_{t-1}, y_{t-2})$, with r_t the amount of rain on day t
- $w \in \mathbb{R}^m$ a vector of unknown parameters
- $\epsilon \sim N(0, \sigma^2)$.

Linear model for rain and flow data

Full regression model (for one river):

$$\mathbf{y} = \mathbf{X}\mathbf{w} + \epsilon$$

- $\mathbf{y} \in \mathbb{R}^n$ where n is the number of days in the dataset
- $\mathbf{w} \in \mathbb{R}^{m \times n}$.

Maximum likelihood and mean squared error:

$$\hat{\mathbf{w}} = \underset{\mathbf{w} \in \mathbb{R}^n}{\operatorname{argmin}} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$$

$$MSE = \frac{1}{n}\|\mathbf{X}\hat{\mathbf{w}} - \mathbf{y}\|^2$$

Functional clustering

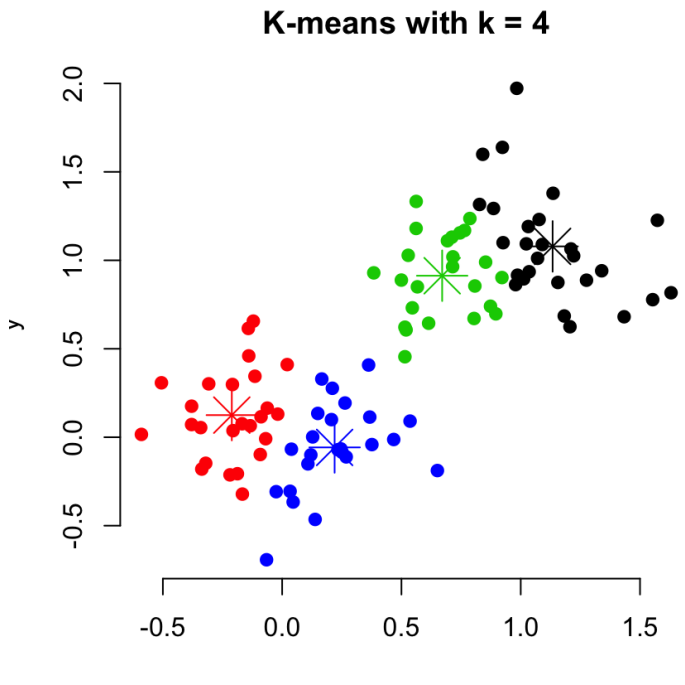
Aim: compare rivers by their fit and prediction accuracy.
Methods:

- Derive a shared model for each pair of rivers. Use the mean squared error of this fitted shared model to derive a similarity metric in order to find similarities between pairs of rivers.
- Produce a clustering algorithm that groups together rivers that perform well, with regard to fit and prediction accuracy, under the same shared model.

k-means clustering

Find k center points, $\mathbf{c_1}, \mathbf{c_2}, \ldots, \mathbf{c_k}$ and mapping $I : \{1, 2, \ldots, n\} \to \{1, 2, \ldots, k\}$ to minimize

$$\sum_{i=1}^{n} \|\mathbf{x_i} - \mathbf{c}_{I(i)}\|^2.$$



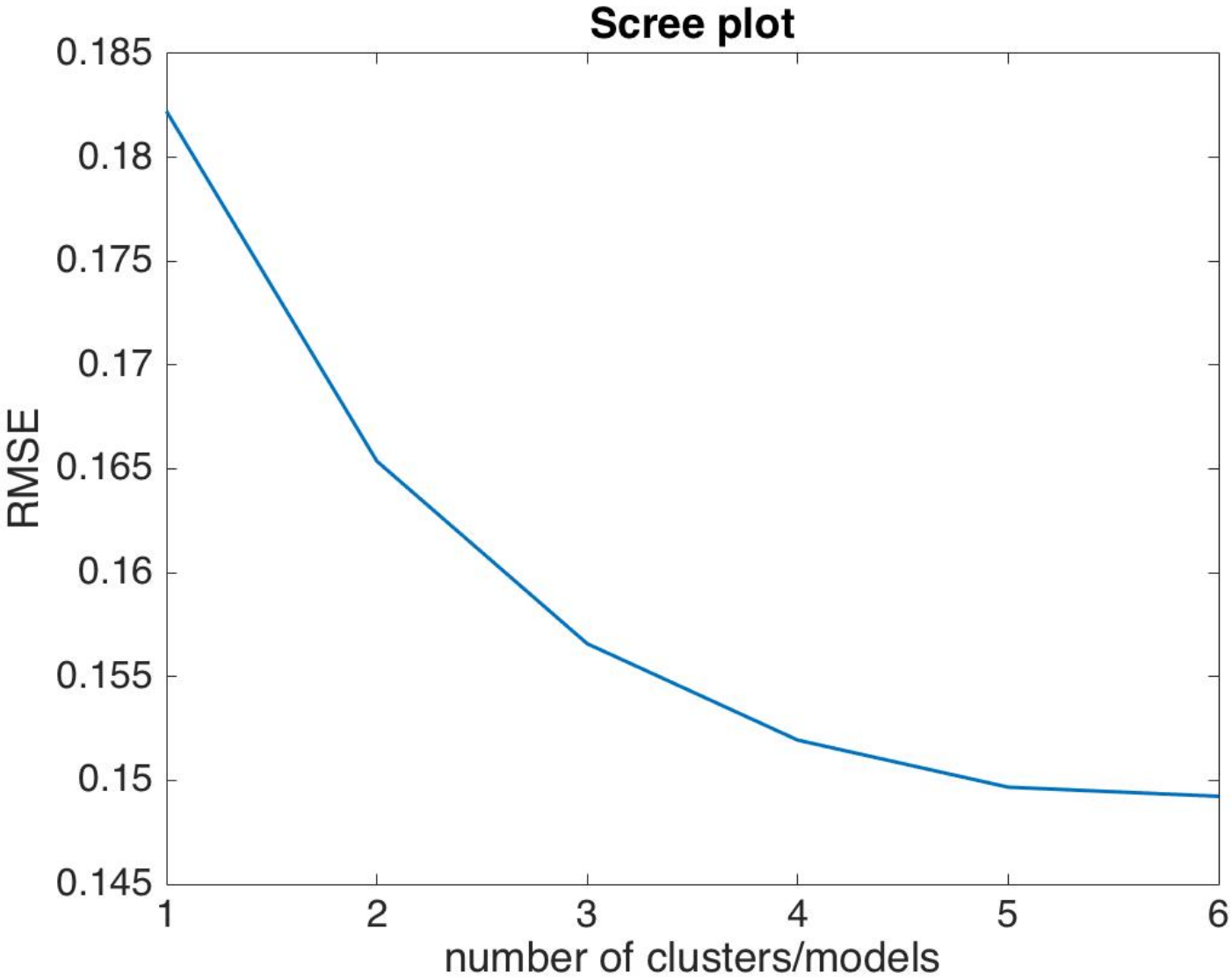
k-riveroid clustering

Find k regression model parameters, $\mathbf{w_1}, \mathbf{w_2}, \ldots, \mathbf{w_k}$ and mapping $I : \{1, 2, \ldots, n\} \to \{1, 2, \ldots, k\}$ to minimize

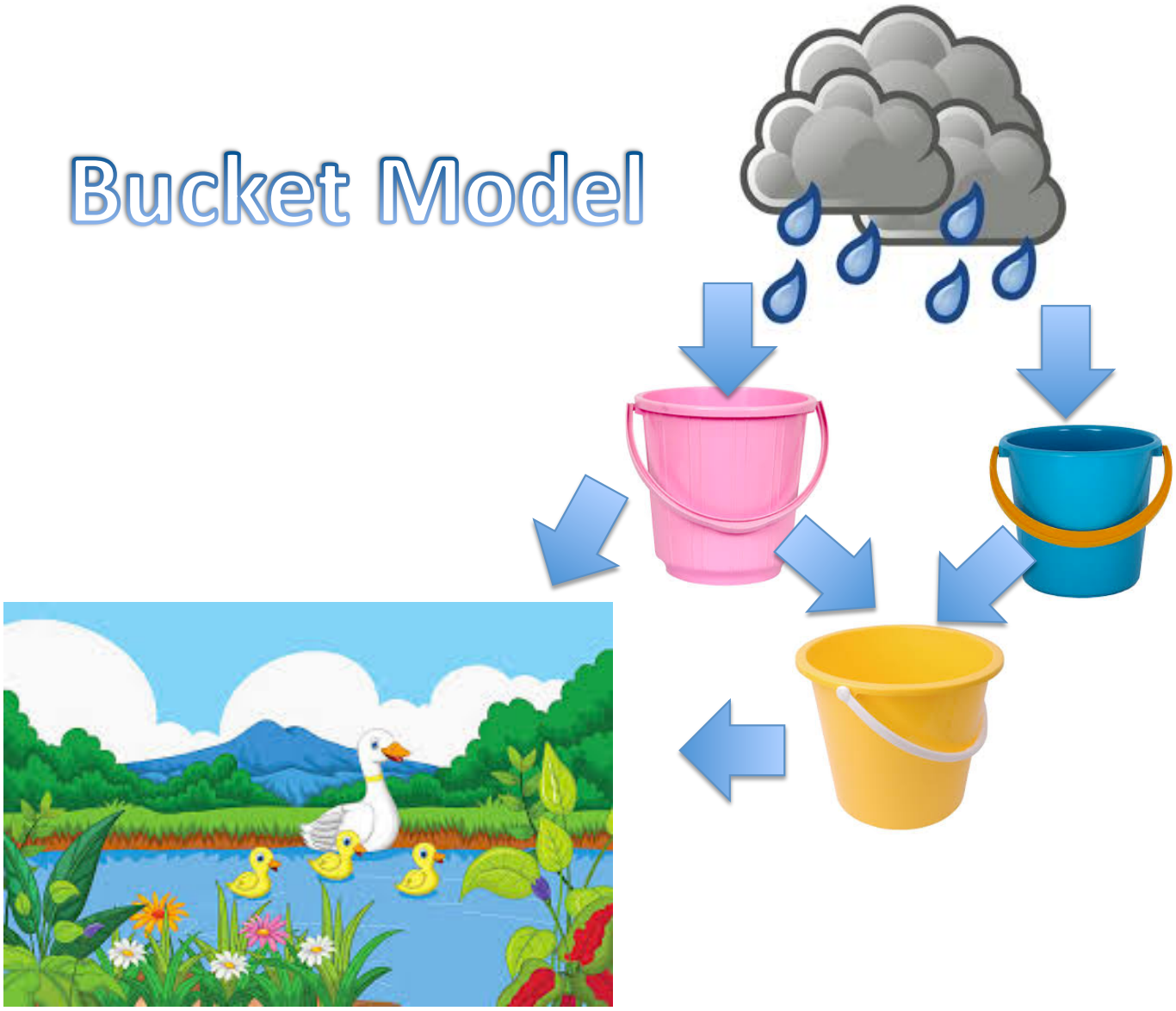
$$\sum_{i=1}^{n} \|\mathbf{X_i}\mathbf{w}_{I(i)} - \mathbf{y_i}\|^2.$$

- Randomly choose k regression vectors.
- For each river, determine which regression vector gives the closest fit to the data.
- Use this information to assign grouping labels of rivers.
- For each group create a shared model and determine the regression vector which maximizes the likelihood.
- Update the regression vectors and repeat until convergence.

k-riveroid clustering



Linear bucket model



Linear Dynamical System bucket model

Let $\mathbf{x}^{(t)}$ be the amount of water in the system of buckets at time t. Then

$$\mathbf{x}^{(t+1)} = M\mathbf{x}^{(t)} + \underline{\alpha} r^{(t)},$$
$$y^{(t)} = \mathbf{c}^{\mathbf{T}}\mathbf{x}^{(t)} + \epsilon.$$

Maximum likelihood Estimate:

$$[\hat{M}, \hat{\underline{\alpha}}, \hat{\mathbf{c}}] = \underset{M,\underline{\alpha},\mathbf{c}}{\operatorname{argmin}} \sum_{t=1}^{n} (y^{(t)} - \mathbf{c}^{\mathbf{T}}\mathbf{x}^{(t)})^2$$

Subject to equality and inequality constraints defined by bucket model architecture.
We can use tools from control theory: system identification, matching transfer functions.

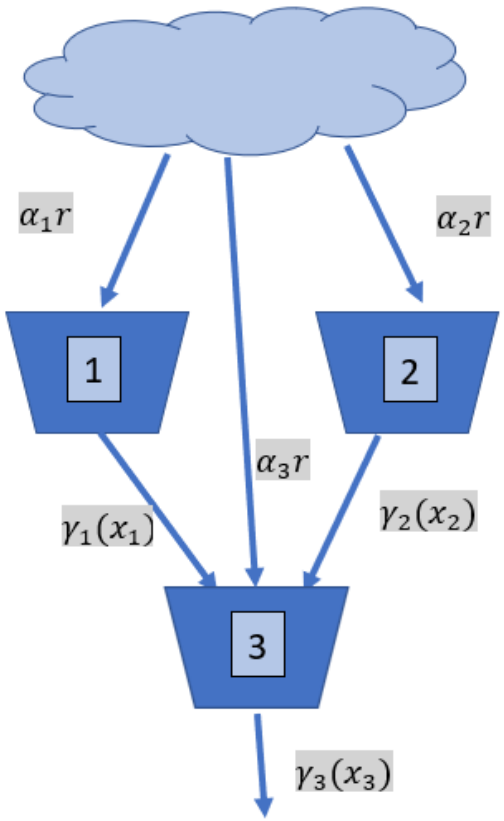
Linear Dynamical System bucket model

$$M = \begin{bmatrix} A_1 & 0 & 0 & 0 \\ B_1 & A_2 & 0 & 0 \\ 0 & B_2 & A3 & 0 \\ 0 & 0 & \ddots & \ddots \end{bmatrix}$$

where $A_k = diag(\delta_j)$, with $\delta_j = 1 - \gamma_{k,j} - \sum_{i=1}^{N-k+1} \beta_{k,j,i}$ and

$$B_k = \begin{bmatrix} \beta_{k,1,1} & \beta_{k,2,1} & \cdots & \beta_{k,N-k+1,1} \\ \beta_{k,1,2} & \beta_{k,2,2} & \cdots & \beta_{k,N-k+1,2} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{k,1,N-k+1} & \beta_{k,2,N-k+1} & \cdots & \beta_{k,N-k+1,N-k+1} \end{bmatrix}.$$

ODE model

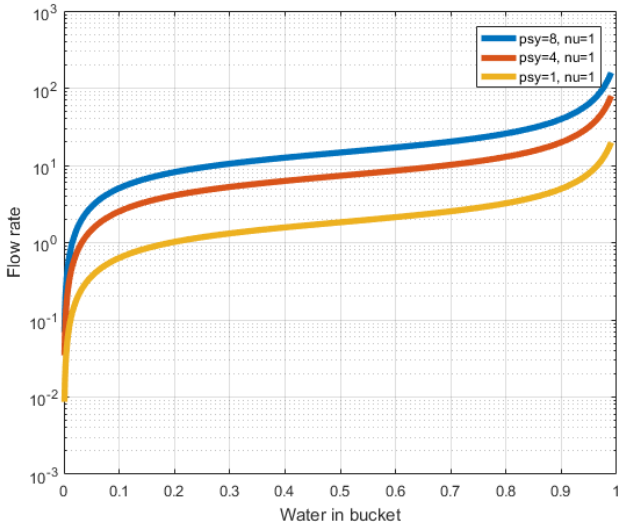


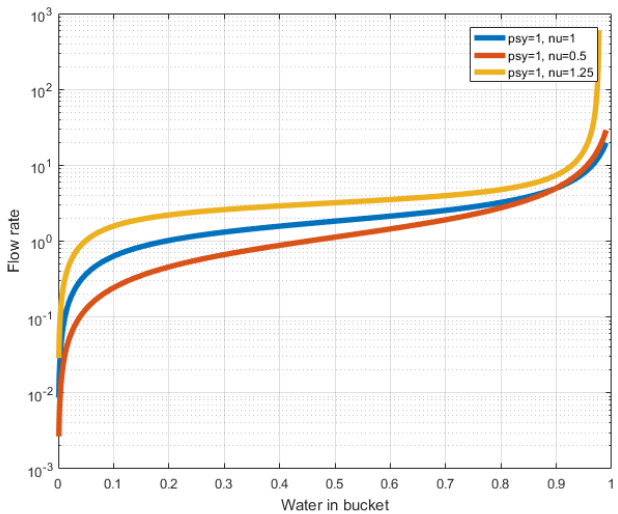
ODE modelling of the system:

$$\dot{x}_1 = \alpha_1 \dot{r} - \gamma_1(x_1)$$
$$\dot{x}_2 = \alpha_2 \dot{r} - \gamma_2(x_2)$$
$$\dot{x}_3 = \alpha_3 \dot{r} - \gamma_3(x_3) + \gamma_1(x_1) + \gamma_2(x_2)$$
$$f = \gamma_3(x_3) + \epsilon$$

where $\gamma_i(x_i)$ is the rate of flow from the bucket i, if bucket i contains x_i units of water.

Method 1: Nonlinear Continuous function





Dripping/Overflowing function γ_i?

$$\gamma = \psi tan\left(\nu\left(\frac{1+\sqrt{5}}{2}\right)^{\frac{1}{\nu}}\frac{\pi}{2}x - \nu\right) - tan(\nu)$$

Where ψ will change the rate of flow for low and high river capacity, x, and ν will change the rate when the river is at high capacity .
We have $x = 1$ as full capacity of the river and the bucket empty at $x = 0$.
We considered the use of a discrete flow γ but would be trickier to use.

MLE: Nonlinear Continuous

Find the Maximum likelihood estimate (MLE)

$$[\hat{\alpha}, \hat{\psi}, \hat{\nu}] = \underset{\hat{\alpha},\hat{\psi},\hat{\nu}}{\text{argmin}} \int_0^t (\gamma_3(x_3) - f(t))^2$$

This is difficult as it requires integrating the nonlinear continuous dynamical system.

Further Research

- More computation using additional data (6 river sites were used previously). This includes adding rainfall (15 minutes as opposed to daily) and adding more river sites.
- Extend the clustering algorithm to the nonlinear system
- More sophisticated bucket structure
- Sensitivity analysis to improve the clustering algorithm
- Continuous prior distribution on model parameters, e.g mixture of Gaussian