Bayesian fitting procedures for hydrological point processes

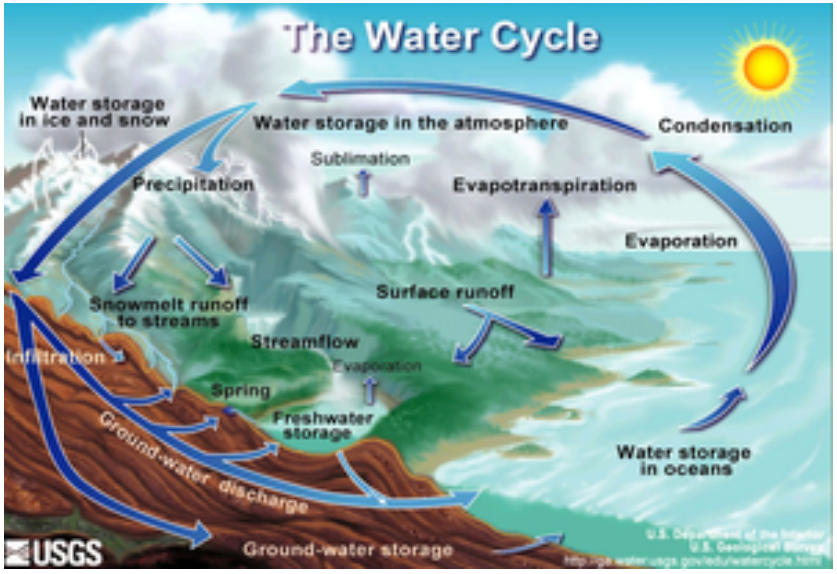
Some people walk in the rain, others just get wet.
(Roger Miller)

Katharina Parry
with assistance by Diane Park and Oliver Hannaford

November 2014

Outline

Hydrologic models.



- Process-based models:
  Try to represent the physical process, e.g. microsimulation of subsurface flows
- Stochastic models:
  Link a certain input (in our case rainfall measurements) to the model output (in our case forecasts of rainfall).

Poisson point cluster model

- We denote arrival times of rainstorms as T_i
- Assume time periods between adjacent rainstorms are Exponential distributed with mean λ^{-1}.
- Storms consist of clusters of rain cells, where for jth cell in the ith storm:
  1. S_{ij} is the arrival time of rain cells, where $S_{ij} - T_i$ are Exponentially distributed with mean β
  2. L_{ij} is the cell lifetime, which are Exponentially distributed with mean η, so that the storm terminates at time $S_{ij} + L_{ij}$.
  3. X_{ij} are random variables representing the cell intensities and are considered to remain constant throughout the lifetime of the cells.

Relationship between variables

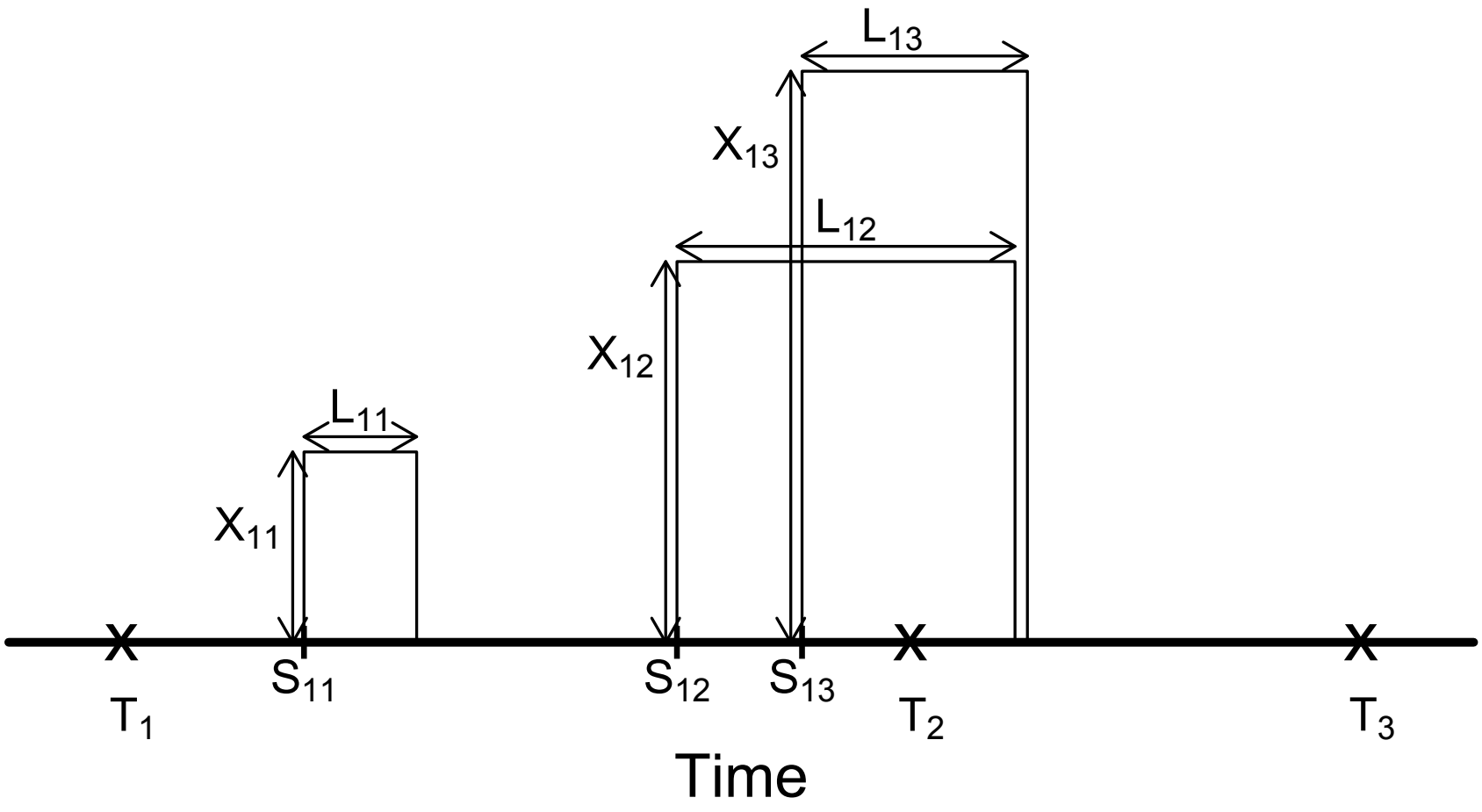


Figure : Random variables $\{(S_{ij}, L_{ij}, X_{ij})\}$ of Poisson cluster model.

Parameters of interest

In summary, the basic model has the following four parameters:

- λ^{-1} the mean time between adjacent storm origins;
- β^{-1} the mean waiting time for a cell origin after a storm origin;
- ν the mean number of rain cells per storm;
- η^{-1} the mean cell lifetime

There is one final parameter, the scale parameter μ, which accounts for the overall level of rainfall in a given rainfall system.

Rainfall measurements

- Rainfall data is usually available in aggregated form, e.g. here measured in amount collected over 5 minute periods
- Sourced from a single site (Kelburn, Wellington) from 1945-2004

- $\rightarrow$ Dealing with a large amount of data

Summary statistics

- Scaleless data properties are used to estimate the model parameters: the autocorrelation, the skewness, the coefficient of variation and the proportion of dry days.
- These summary statistics are calculated for the data at various levels of aggregation.
- In particular, we worked with the following summary statistics:
  cv10m, ac10m, sk10m, cv1h, ac1h, sk1h,
  cv6h, ac6h, sk6h, cv24h, ac24h, sk24h, pd24h

Why not MCMC?

- Complex models $\rightarrow$ intractable likelihoods.
- However, sampling from the posterior using conventional MCMC methods is still computationally expensive
- Can use ABC instead in cases where it is possible to simulate data in a reasonable amount of time.

Idea of ABC

Consider a rainfall model involving the set of unknown parameters denoted as θ.

Standard Bayesian inference Specify likelihood $\pi(\boldsymbol{y}|\theta)$ and prior $\pi(\theta)$. Multiplication gives us the posterior $\pi(\theta|\boldsymbol{y})$.

Approximate Bayesian inference No exact form of $\pi(\theta|\boldsymbol{y})$ is calculated. It is reconstructed using the observed summary statistics, $\boldsymbol{s}$, derived from the original data. Essentially, the observed summary statistics are used to replace the original observations.

Details

In other words:

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) \approx \quad \pi(\boldsymbol{\theta}|\boldsymbol{s}) \propto \pi(\boldsymbol{s}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$$

where the approximation of the likelihood is defined as

$$\pi(\boldsymbol{s}|\boldsymbol{\theta}) = \int \pi(\boldsymbol{y}_{sim}|\boldsymbol{\theta}) K_h\Big(S(\boldsymbol{y}_{sim}) - \boldsymbol{s}\Big) d\boldsymbol{y} = \mathbb{E}_{\boldsymbol{\theta}}[K_h(\boldsymbol{s}_{sim} - \boldsymbol{s})],$$

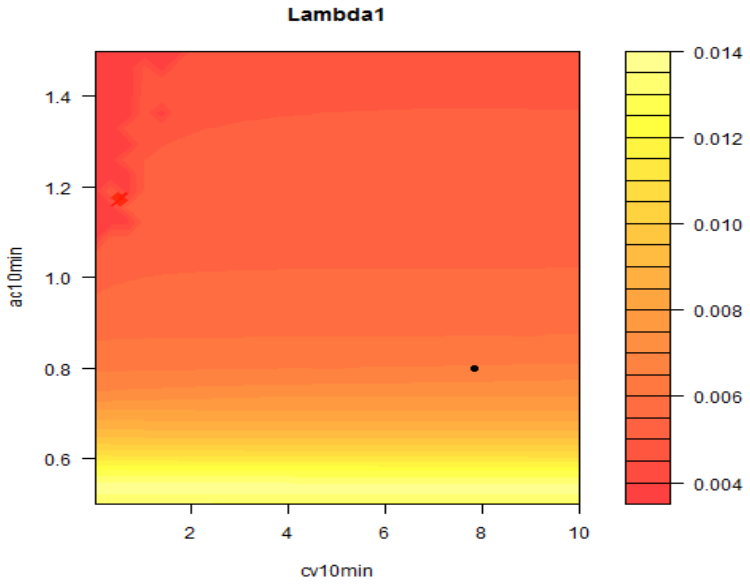
where $K_h(\cdot)$ is a d-dimensional kernel density, $\boldsymbol{y}_{sim}$ is simulated data and the parameter h is a non-negative measure of the width of the kernel.
As the parameter h approaches 0, the approximate likelihood converges towards the true likelihood $p(\boldsymbol{y}|\boldsymbol{\theta})$

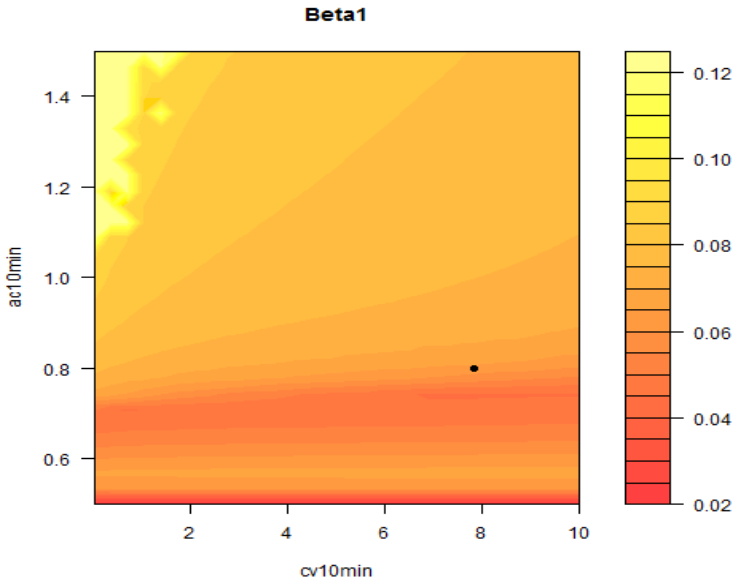
Curse of Dimensionality

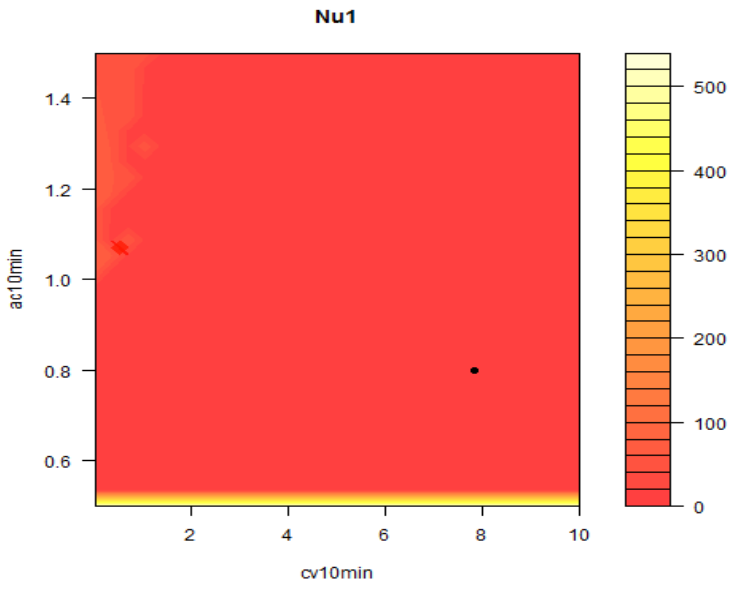
- ABC works for a **well-defined model** and **sufficient statistics**.
- However, ABC suffers from the curse of dimensionality.
- Review of ABC literature shows that examples with only 4 dimensions or less.
- Using the ABC algorithm with 13 summary statistics is not advised.
- $\rightarrow$ Need to reduce the number of summary statistics.

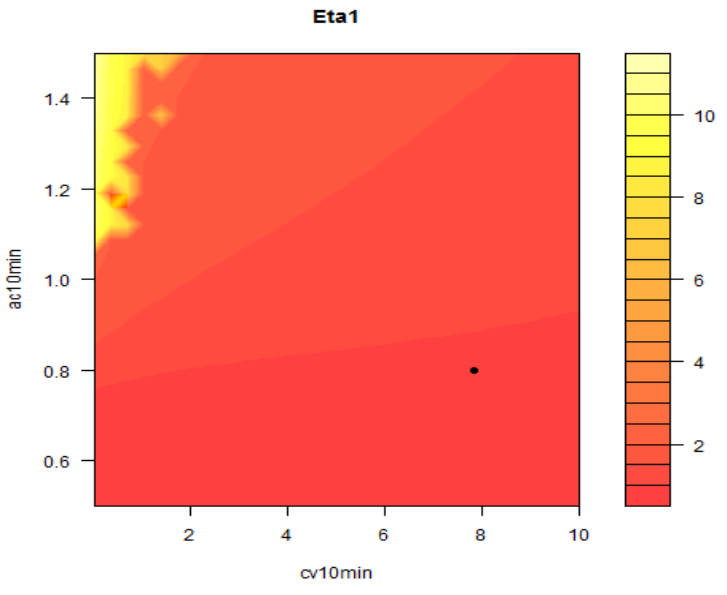
Question:
Which summary statistics are particularly useful
in the estimation process?

Contour plots

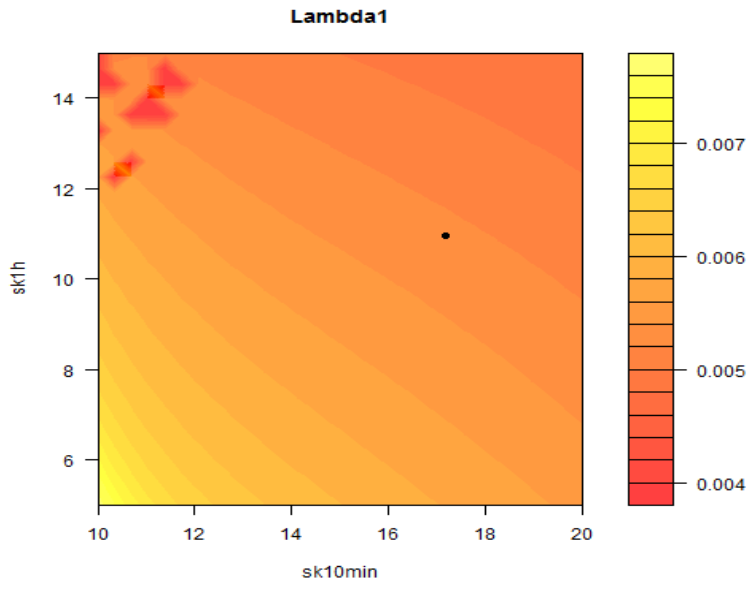


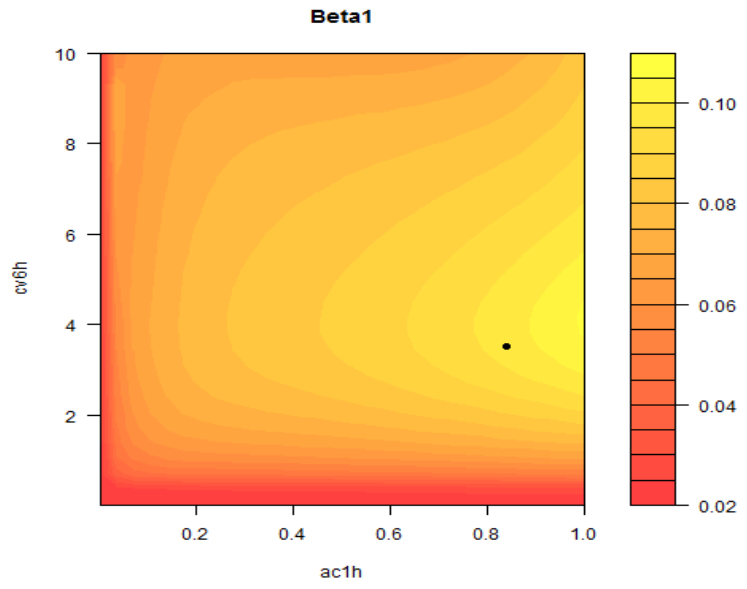


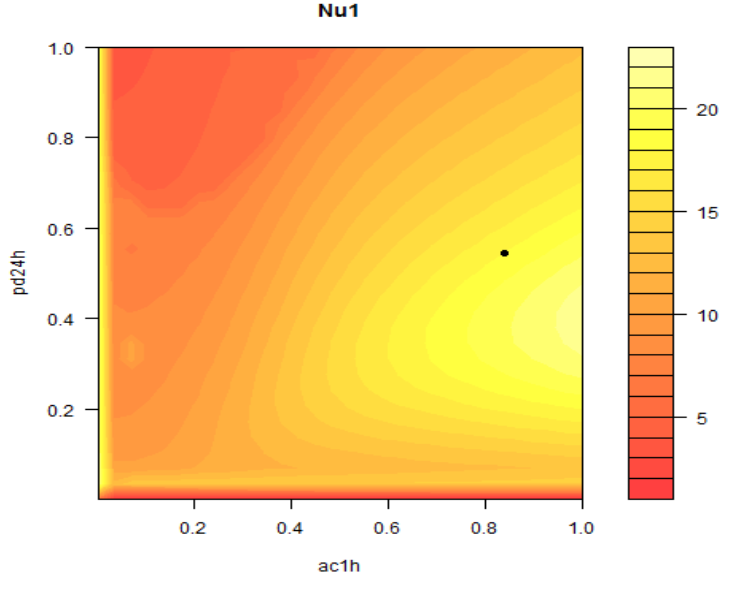


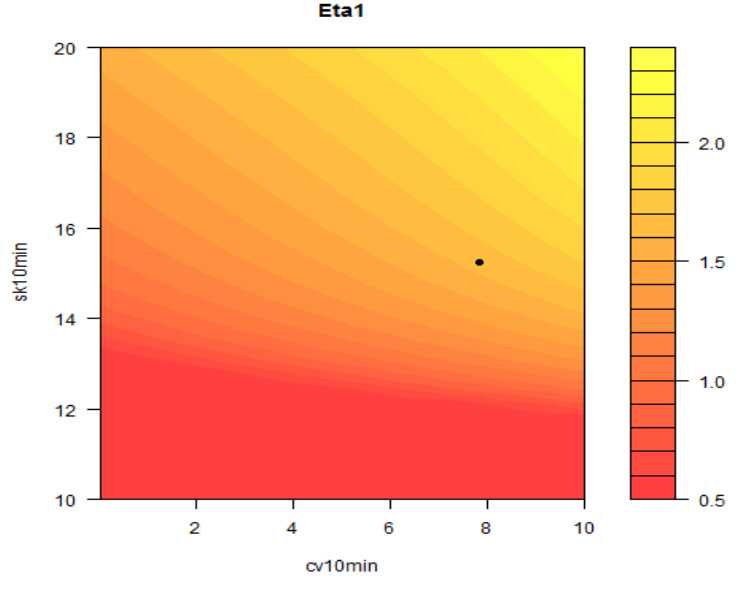


More contour plots

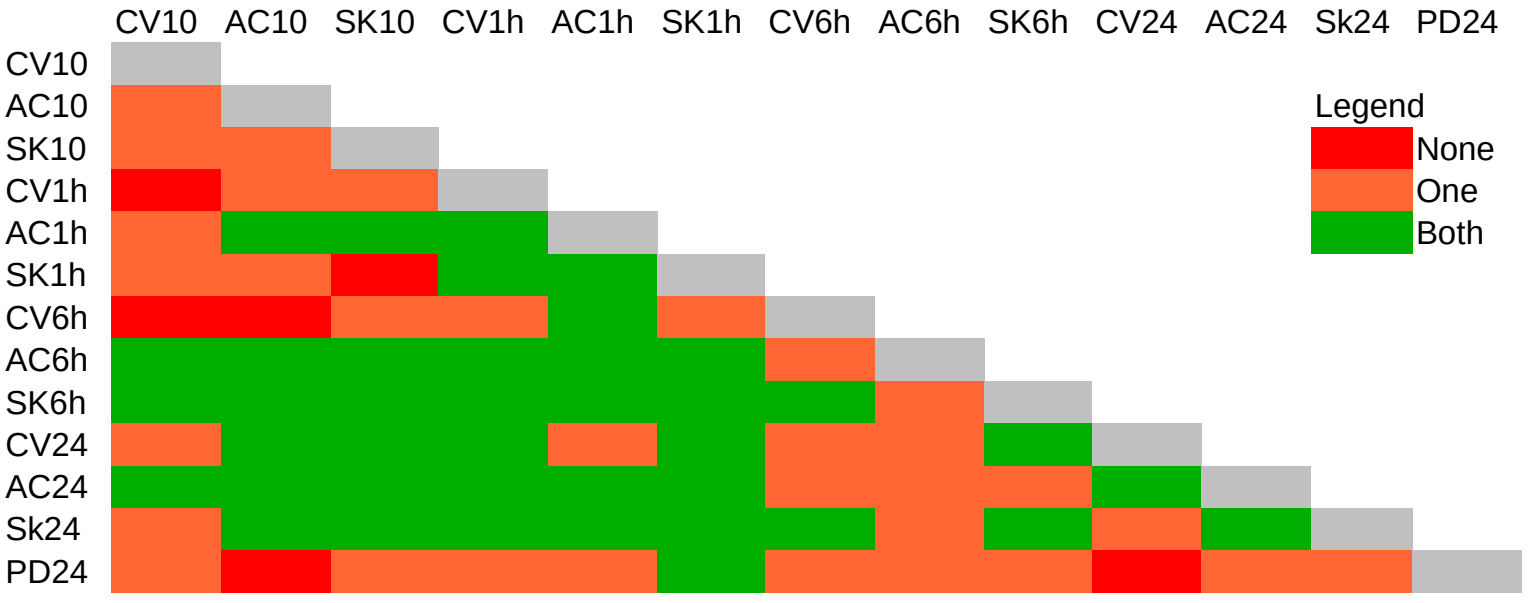




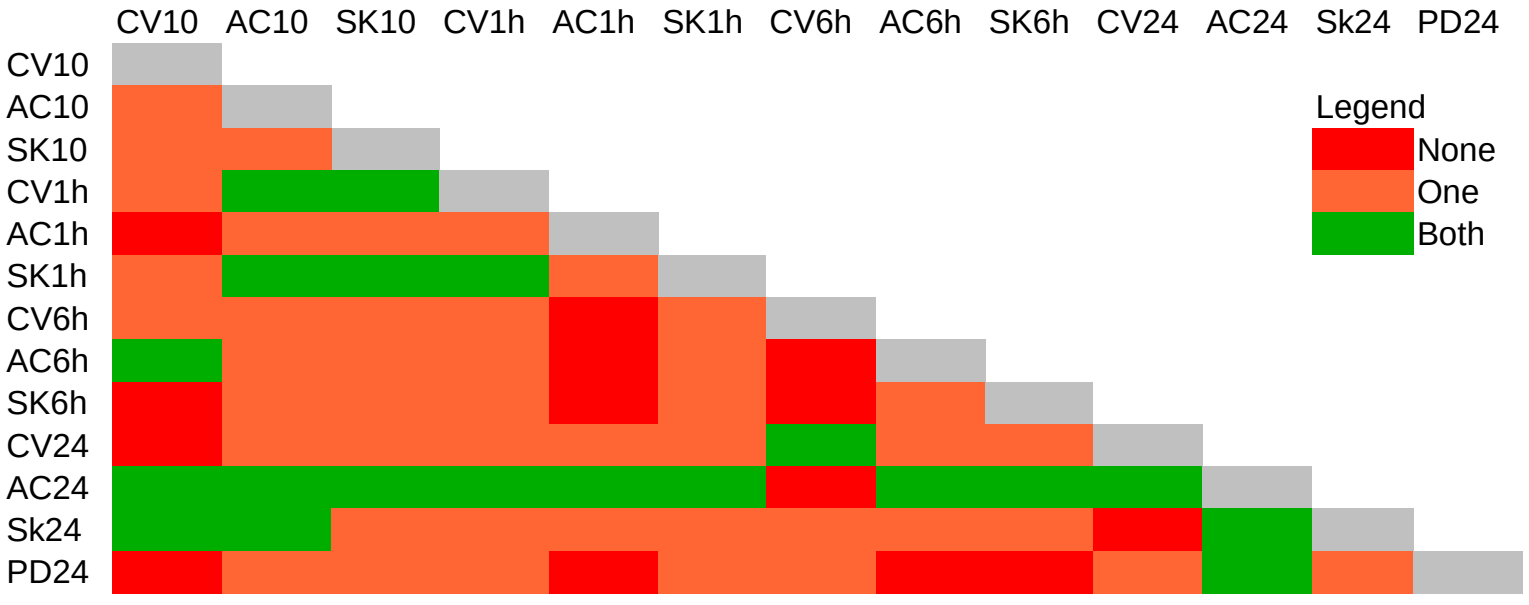




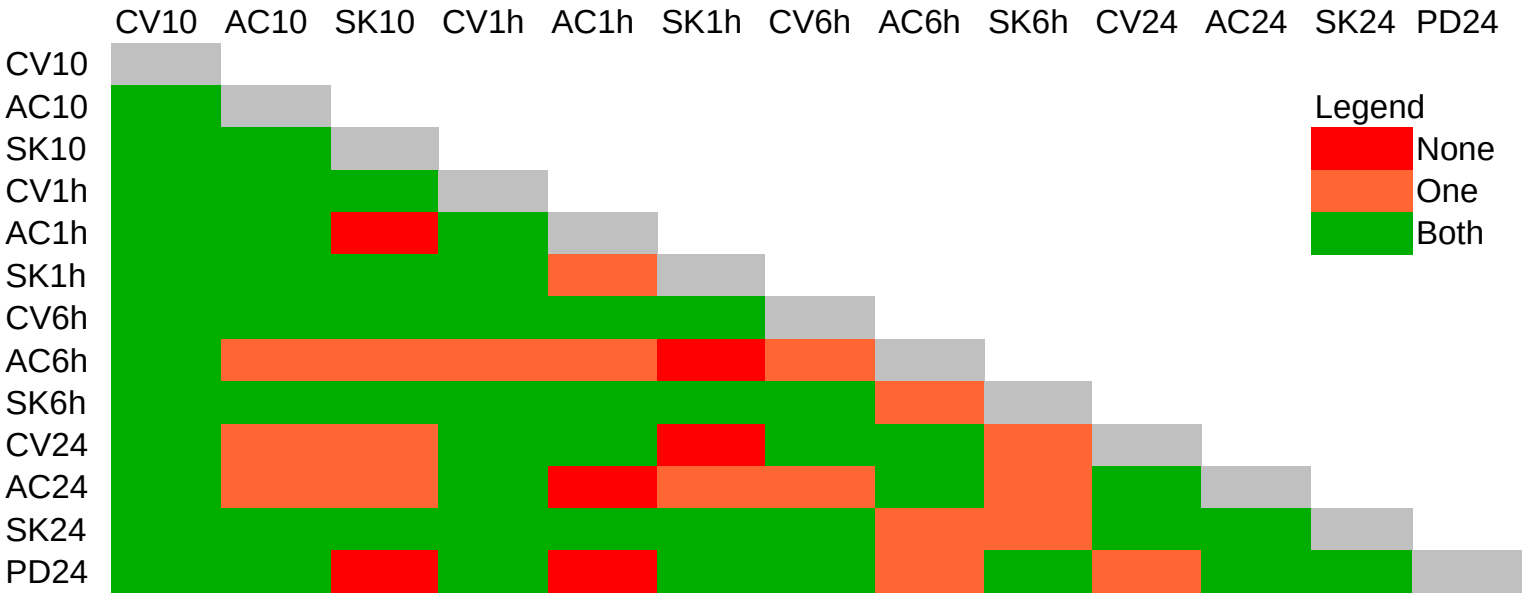
Matrix of bi-variable relationships in estimates for λ



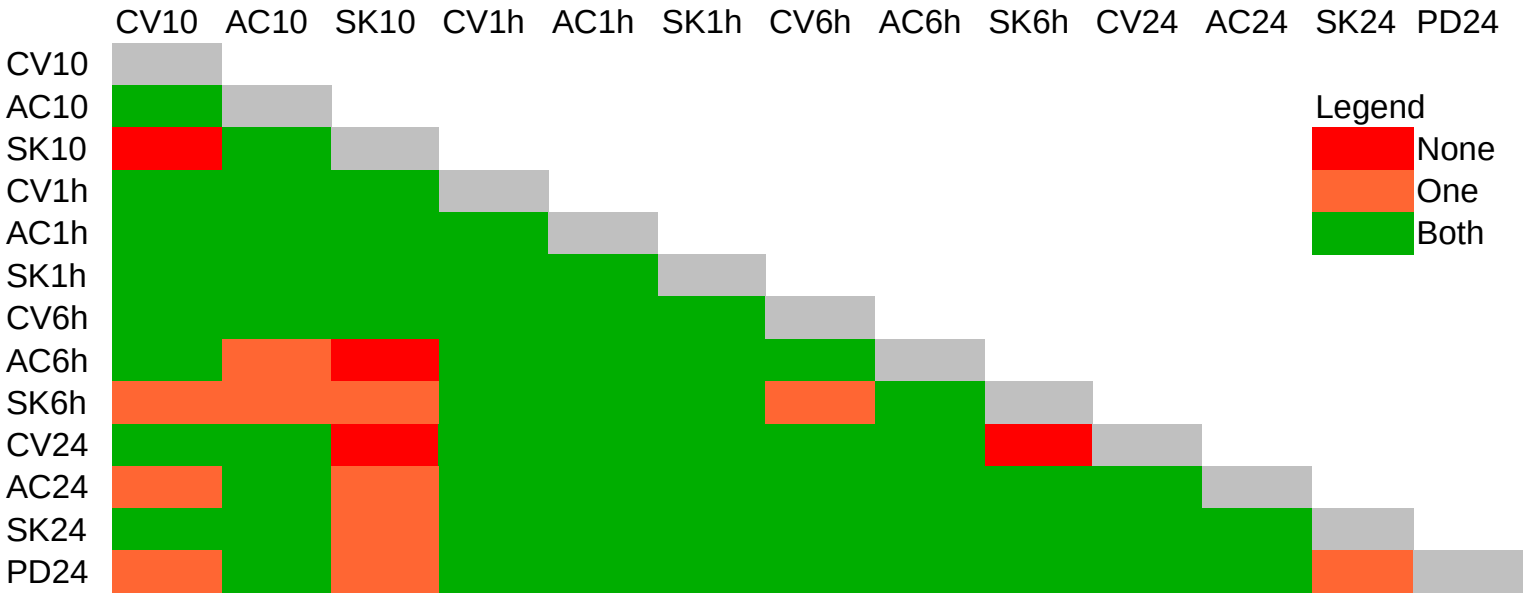
Matrix of bi-variable relationships in estimates for β



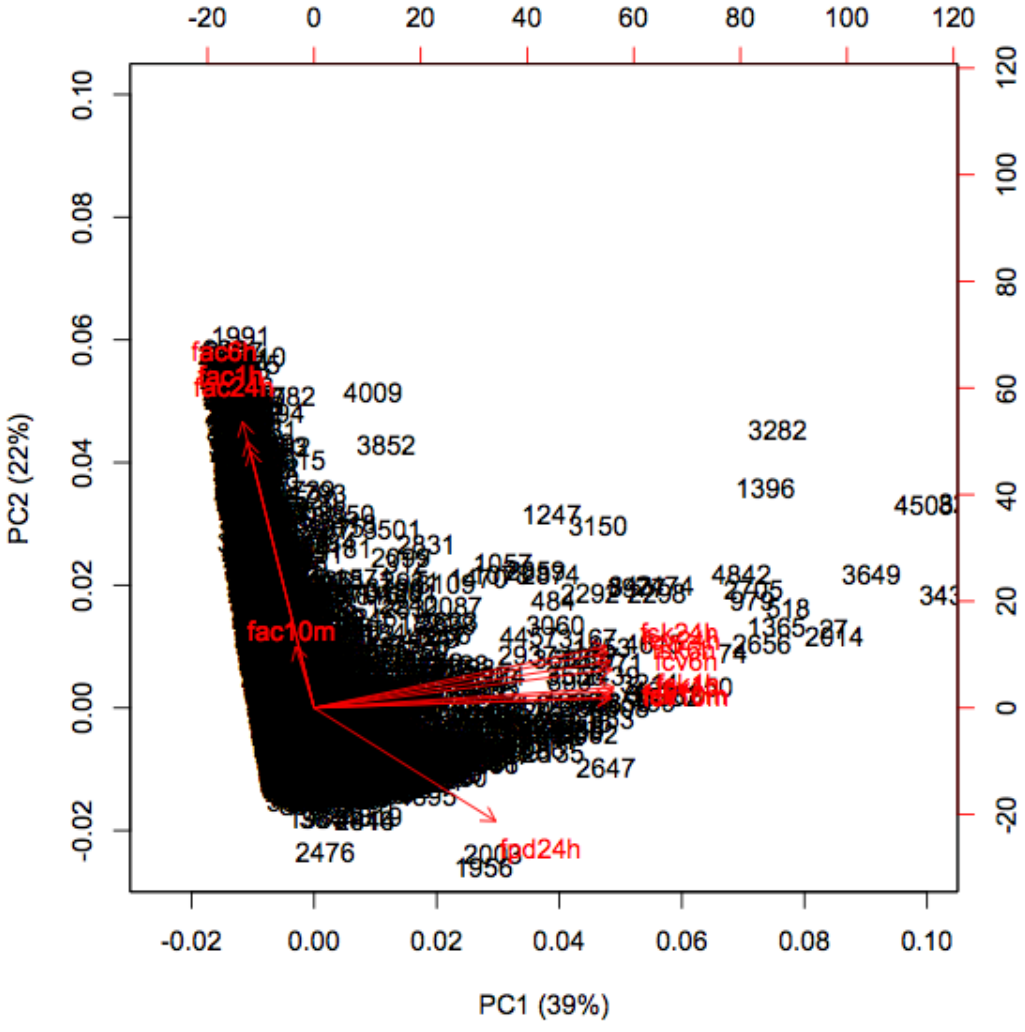
Matrix of bi-variable relationships in estimates for ν



Matrix of bi-variable relationships in estimates for η



Biplot from PCA of matrix of simulated summary statistics



Goals

- Redo ABC analysis with only three summary statistics: AC24h, PD24h and maybe SK1h
- Any suggestions?

THANKS FOR LISTENING!