Model based Clustering of Multiple Time Series

Presentation by C. Simon
Paper by Sylvia Fruhwirth-Schnatter, and Sylvia Kaufmann
Published in Journal of Business and Economic Statistics

Applied Math and Statistics, UCSC

March 9, 2010
Outline

- Notation
- Model Based Clustering
- Bayesian Estimation
- Unknown Number of Clusters
- Future Work
Model Based Clustering of Multiple Time Series

- Data
  \[ \{y_{i,t}\} \quad t = 1, \ldots, T, \quad i = 1, \ldots, N \]
  - $T$ number of time steps
  - $N$ number of units
Data

\[ \{y_{i,t}\} \quad t = 1, \ldots, T, \quad i = 1, \ldots, N \]

- \(T\) number of time steps
- \(N\) number of units

Usually data is pooled, that is a single parameter is inferred for all time series.
Suppose the goal is forecast 

\[ y_{it} = c_k + \delta_{1,k}y_{i,t-1} + \cdots + \delta_{p,k}y_{i,t-p} + \varepsilon_{i,t} \]

\[ \varepsilon_{i,t} \sim \mathcal{N}(0, \sigma_k^2) \]

\[ k = 1, \ldots, K \]

\( K \) is the number of groups.
Model Formulation and Notation

\{y_{i,t}\} \quad t = 1, \ldots, T, \quad i = 1, \ldots, N

\mathbf{y}_i = \{y_{i1}, \ldots, y_{iT}\}

\theta \text{ is unknown}

\begin{align*}
p(\mathbf{y}_i | \theta) &= \prod_{t=1}^{T} p(y_{it} | \mathbf{y}_{i}^{t-1}, \theta) \\
\mathbf{y}_{i}^{t-1} \text{ are observations up to time } t - 1
\end{align*}
Typically,

\[ y_{it} | y_{i}^{t-1}, \theta \sim N(\hat{y}_{it|t-1}(\theta), C_{it|t-1}(\theta)) \]

- So far we have implicitly assumed that all models are the same but with different parameters.
- the joint is therefore

\[
p(y_1, ..., y_N | S_1, ..., S_N, \theta_1, ..., \theta_K) = \prod_{k=1}^{K} \prod_{\{S_i = k\}} p(y_i | \theta_k)
\]
Assuming complete prior ignorance about group membership

\[ \Pr(S_i = k | \eta_1, ..., \eta_K) = \eta_k \]

\[ \eta_K = 1 - \sum_{k=1}^{K} \eta_k \]

\( \eta_k = \) relative size of group \( k \).
they find the normal to be too restrictive so they use a $t_\nu$ and represent it with a mixture of normals

$$y_{it|\mathbf{y}_{i}^{t-1}, \lambda_i, \theta} \sim N(\hat{y}_{it|t-1}(\theta), \frac{C_{it|t-1}(\theta)}{\lambda_i})$$

where

$$\hat{y}_{it|t-1}(\theta) = c_k + \delta_{1,k}y_{i,t-1} + \cdots + \delta_{p,k}y_{i,t-p}$$

and

$$C_{it|t-1}(\theta) = \sigma^2$$
Model Based Clustering

- $N$ time series in $K$ groups.
- $S_i$ is the unit specific parameter

\[
p(y_i | S_i, \theta_1, \ldots, \theta_K) = p(y_i | \theta_{S_i})
\]

\[
p(y_i | \theta_{S_i}) = p(y_i | \theta_k) \text{ if } S_i = k
\]
Bayesian Estimation

Many different models used in time series

\[ p(\theta_k) \] depends on model. For conjugacy, usually \textit{Normal} for mean and \textit{Gamma} for \( \sigma^{-2} \)

set \( \phi = (\eta_1, \ldots, \eta_K) \),

\[ \phi \sim Dirichlet(e_0, \ldots, e_0) \]

The \textit{Dirichlet} allows for straightforward estimation in MCMC.
Iterate between two steps:

(a) *Classification for fixed parameter*
Sample from

\[ \Pr(S_i = k|y, \theta_1, ..., \theta_K, \phi) \quad k = 1, ..., K \]

(b) *Estimation for a fixed classification*
Known \( S = (S_1, ..., S_k) \),
sample from

\[ p(\theta_1, ..., \theta_K | S, y) \]

and

\[ p(\phi | S, y) \]
Once the model has been identified, it is possible to classify the time series into the various groups by estimating for each time series the posterior classification probability (from MCMC draws),

\[
\Pr(S_i = k | y) = \frac{1}{M} \sum_{m=1}^{M} I\{S_i(m) = k\}
\]

\(M\) is the number of samples.
Unknown number of Clusters

\[ p(y|M_k) = \int P(y_1, \ldots, y_N | \psi, k) p(\psi) d\psi \]

\[ \psi = (\theta_1, \ldots, \theta_K, \phi, S, \lambda) \]

\[ \Rightarrow \Pr(M_k|y) \propto p(y|M_k) \Pr(M_k) \]

\[ \{1, \ldots, K_{max}\} \quad \Pr(M_k) = \frac{1}{K_{max}} \]

\[ \Rightarrow \text{reduces to choosing largest } p(y|M_k) \]
Futurework

- Reversible Jump MCMC
  Generalizes model selection in a more satisfactory way.

- Missing Data
  Data which have different time step sizes, or missing a few data points at some place in the time series.

- Partition is time
  By clustering the units (space) they organize the information in the $K$ clusters as if it the models did not change over time. Model switching at some time say $t^*_i$, will generalize.

- Gaussian Process
  Set up is ripe for giving it a Gaussian process prior.