Analysis of multivariate time series with missing data

Karl Øyvind Mikalsen

Aug, 2019

Outline

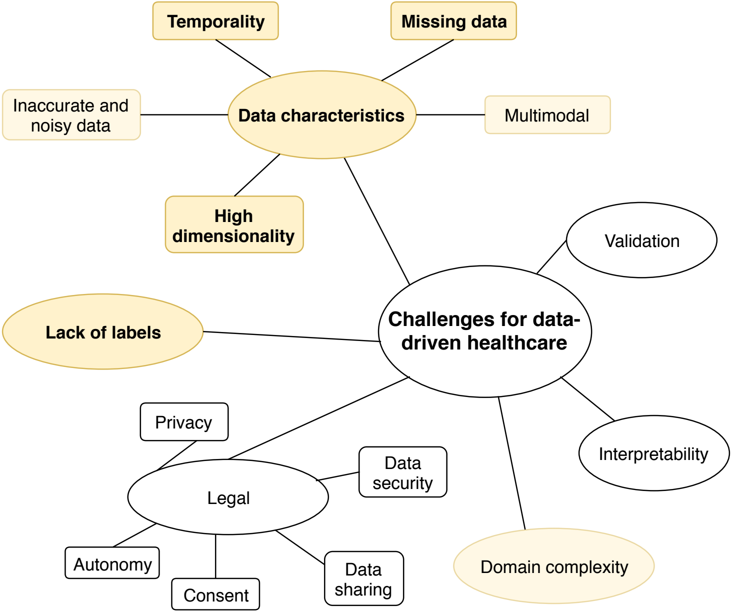
1. Introduction
2. Time series cluster kernel for learning similarities between multivariate time series with missing data
3. Time series cluster kernels to exploit informative missingness and incomplete label information

Outline

1. Introduction

2. Time series cluster kernel for learning similarities between multivariate time series with missing data

3. Time series cluster kernels to exploit informative missingness and incomplete label information


Temporality
Missing data
Inaccurate and noisy data
Data characteristics
Multimodal
High dimensionality
Validation
Lack of labels
Challenges for data-driven healthcare
Privacy
Data security
Interpretability
Legal
Autonomy
Consent
Data sharing
Domain complexity

Outline

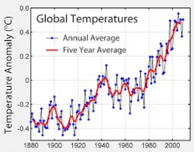
1. Introduction
2. Time series cluster kernel for learning similarities between multivariate time series with missing data
3. Time series cluster kernels to exploit informative missingness and incomplete label information

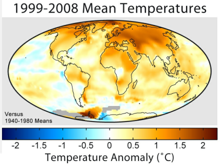
Paper I

K. Ø. Mikalsen, F. M. Bianchi, C. Soguero-Ruiz and R. Jenssen, "**Time series cluster kernel for learning similarities between multivariate time series with missing data**", *Pattern Recognition*, Apr. 2018, Vol. 76, pp 569–581, doi: https://doi.org/10.1016/j.patcog.2017.11.030.

Time series analysis

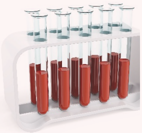
Climate studies



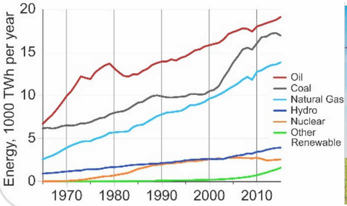


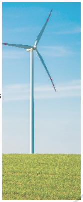
Finance

Medicine



Energy consumption





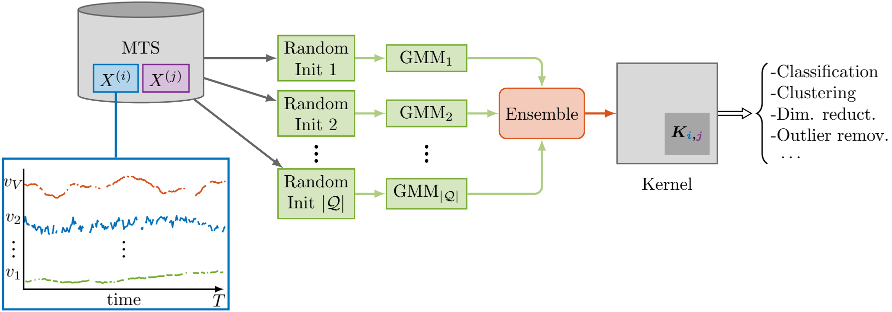
Objective

Create a **kernel method** for **multivariate** time-series with **missing** data which is **robust** to hyperparameters.

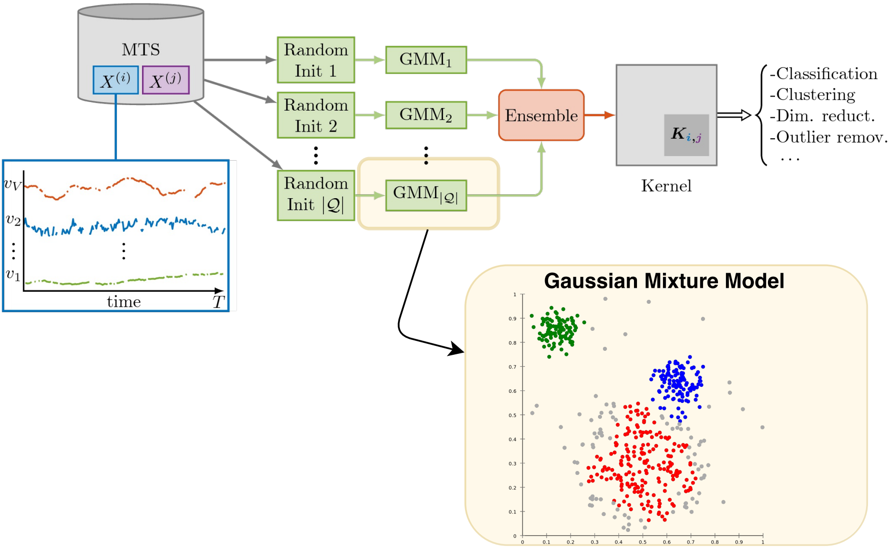
Solution

1. Probabilistic formulation.
   - Can deal with missing data effectively.
   - Naturally extended to multivariate data.
2. Ensemble learning.
   - Robustness to hyperparameters.

Time series Cluster Kernel (TCK)

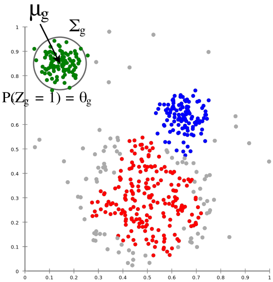


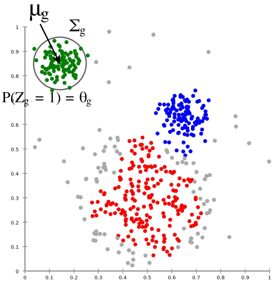
Time series Cluster Kernel (TCK)



Ordinary GMM

- Mixture of G normally distributed components.
- Described by the mixing coefficients θ_g, means μ_g and covariances Σ_g.





Ordinary GMM

- Mixture of G normally distributed components.
- Described by the mixing coefficients θ_g, means μ_g and covariances Σ_g.

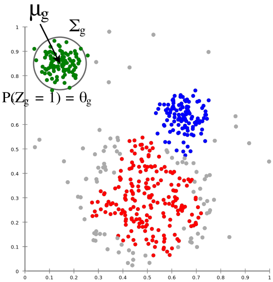
GMM for multivariate time-series with missing data

- Time-dependent means $\mu_g = \{\mu_{gv} \in \mathbb{R}^T \mid v = 1, ..., V\}$.
- Diagonal covariance $\Sigma_g = diag\{\sigma_{g1}^2, ..., \sigma_{gV}^2\}$, constant over time.
- *Missing at random* assumption.

Bayesian GMM for MTS with missing data

Ordinary GMM

- Mixture of G normally distributed components.
- Described by the mixing coefficients θ_g, means μ_g and covariances Σ_g.



GMM for multivariate time-series with missing data

- Time-dependent means $\mu_g = \{\mu_{gv} \in \mathbb{R}^T \mid v = 1, ..., V\}$.
- Diagonal covariance $\Sigma_g = diag\{\sigma_{g1}^2, ..., \sigma_{gV}^2\}$, constant over time.
- *Missing at random* assumption.

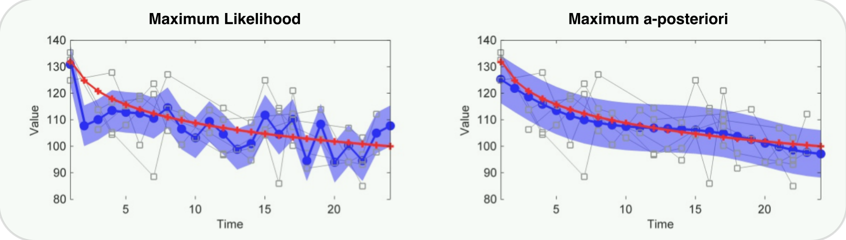
Posterior: $\pi_g = \dfrac{\theta_g \prod_{v=1}^{V} \prod_{t=1}^{T} \mathcal{N}\big(x_v(t) \mid \mu_{gv}(t), \sigma_{gv}\big)^{r_v(t)}}{\sum_{g=1}^{G} \theta_g \prod_{v=1}^{V} \prod_{t=1}^{T} \mathcal{N}\big(x_v(t) \mid \mu_{gv}(t), \sigma_{gv}\big)^{r_v(t)}}.$

Estimation of model parameters $\Theta = \{\theta_g,\ \mu_g,\ \Sigma_g\}$

- **Maximum likelihood:** $\hat{\Theta}_{\mathrm{ML}} = \mathrm{argmax}_{\Theta}\ p(X \mid \Theta)$
  No closed form solution $\implies$ **Expectation Maximization** (EM).
- Problem: Missing data.
- Solution ([Marlin et al, 2012]): **Bayesian approach**, put priors on parameters, $p(\Theta)$,
  and use **Maximum a posteriori** EM: $\hat{\Theta}_{\mathrm{MAP}} = \mathrm{argmax}_{\Theta}\ p(X \mid \Theta)p(\Theta)$.

Estimation of model parameters $\Theta = \{\theta_g,\ \mu_g,\ \Sigma_g\}$

- **Maximum likelihood:** $\hat{\Theta}_{\mathrm{ML}} = \mathrm{argmax}_{\Theta}\, p(X \mid \Theta)$
  No closed form solution $\implies$ **Expectation Maximization** (EM).
- Problem: Missing data.
- Solution ([Marlin et al, 2012]): **Bayesian approach**, put priors on parameters, $p(\Theta)$,
  and use **Maximum a posteriori** EM: $\hat{\Theta}_{\mathrm{MAP}} = \mathrm{argmax}_{\Theta}\, p(X \mid \Theta)p(\Theta)$.



1. Smooth cluster means.
2. Parameters similar to overall mean and covariance for clusters containing few time series.

Informative prior distributions for μ and Σ

Kernel-based Gaussian prior for the mean

$$P(\mu_{gv}) = \mathcal{N}\left(\mu_{gv} \mid m_v,\ S_v\right)$$

m_v empirical mean (for attribute v).
$S_v = s_v\mathcal{K}$, prior covariance matrix.
s_v empirical standard deviation
$\mathcal{K}$ kernel matrix,

$$\mathcal{K}_{tt'} = b_0 \exp(-a_0(t-t')^2), \quad t,\ t' = 1, \ldots, T.$$

Informative prior distributions for μ and Σ

Kernel-based Gaussian prior for the mean

$$P(\mu_{gv}) = \mathcal{N}\left(\mu_{gv} \mid m_v,\ S_v\right)$$

m_v empirical mean (for attribute v).
$S_v = s_v \mathcal{K}$, prior covariance matrix.
s_v empirical standard deviation
$\mathcal{K}$ kernel matrix,

$$\mathcal{K}_{tt'} = b_0 \exp(-a_0(t - t')^2), \quad t,\ t' = 1, \ldots, T.$$

Inverse Gamma distribution prior is for standard deviation

$$P(\sigma_{gv}) \propto \sigma_{gv}^{-N_0} \exp\left(-\frac{N_0 s_v}{2\sigma_{gv}^2}\right)$$

a_0, b_0 and N_0 are user-defined hyperparameters.

Algorithm 1 MAP-EM for DiagGMM

Input $\{(X^{(n)}, R^{(n)})\}_{n=1}^{N}$, Ω and number of mixtures G.

1: Initialize the parameters Θ.

2: E-step. For each MTS $X^{(n)}$, evaluate the posterior probabilities using current parameter estimates,

$$\pi_g^{(n)} = P(Z_g^{(n)} = 1 \mid X^{(n)},\ R^{(n)},\ \Theta).$$

3: M-step. Update parameters using the current posteriors

$$\theta_g = N^{-1} \textstyle\sum_{n=1}^{N} \pi_g^{(n)}$$

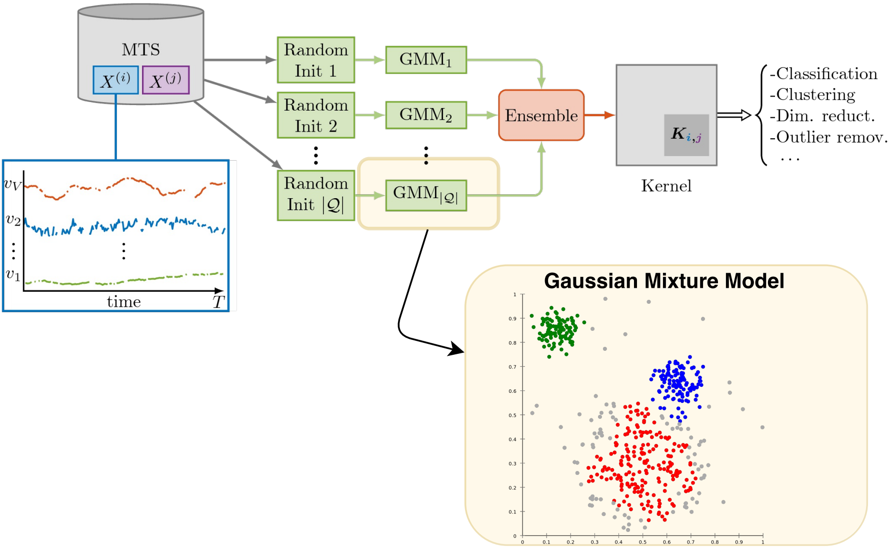
$$\sigma_{gv}^2 = \frac{N_0 s_v^2 + \sum_{n=1}^{N} \sum_{t=1}^{T} r_v^{(n)}(t)\ \pi_g^{(n)} \left(x_v^{(n)}(t) - \mu_{gv}(t)\right)^2}{N_0 + \sum_{n=1}^{N} \sum_{t=1}^{T} r_v^{(n)}(t)\ \pi_g^{(n)}}$$

$$\mu_{gv} = \frac{S_v^{-1} m_v + \sigma_{gv}^{-2} \sum_{n=1}^{N} \pi_g^{(n)} \mathrm{diag}(r_v^{(n)})\ x_v^{(n)}}{S_v^{-1} + \sigma_{gv}^{-2} \sum_{n=1}^{N} \pi_g^{(n)} \mathrm{diag}(r_v^{(n)})}$$

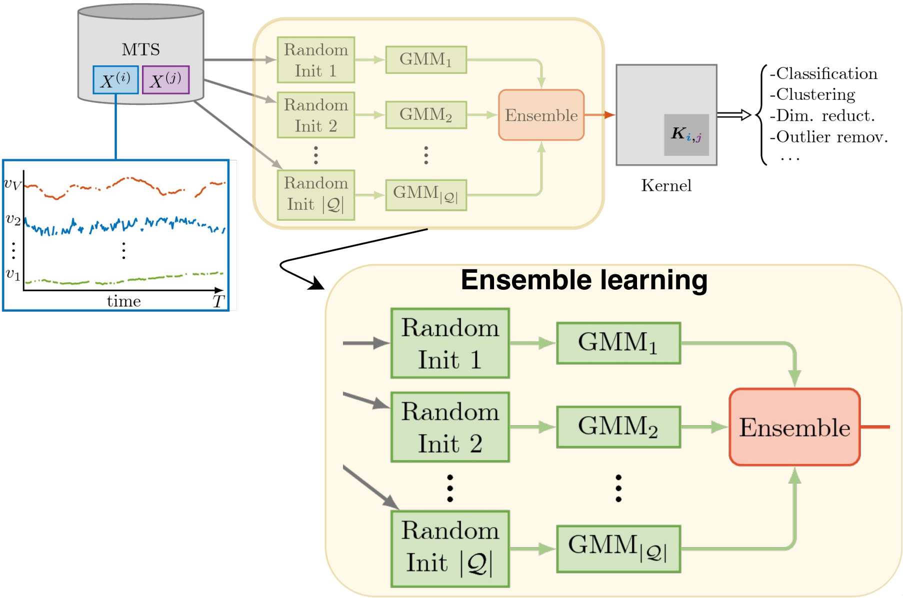
4: Repeat step 2-3 until convergence.

Output Posteriors $\Pi^{(n)} \equiv (\pi_1^{(n)}, \ldots, \pi_G^{(n)})^T$ and mixture parameters Θ.

First part of the TCK: Probabilistic model

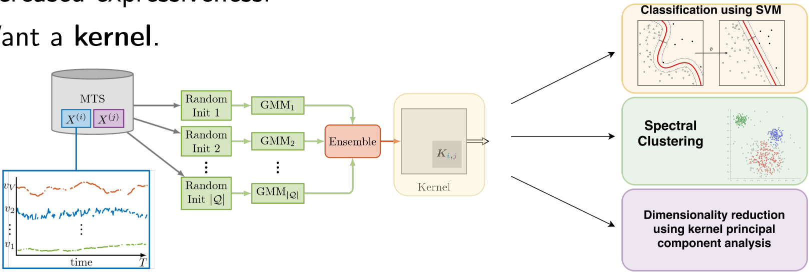


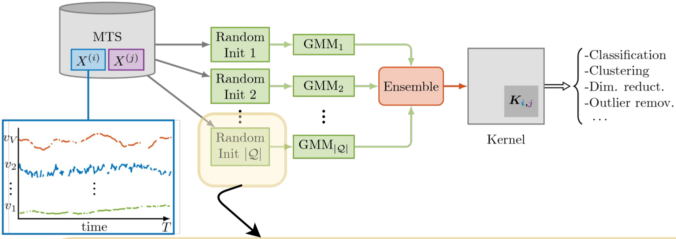
Second part of the TCK: Ensemble learning



Why ensemble learning?

1. Want **robustness** to hyperparameters.
2. Increased expressiveness.
3. Want a **kernel**.

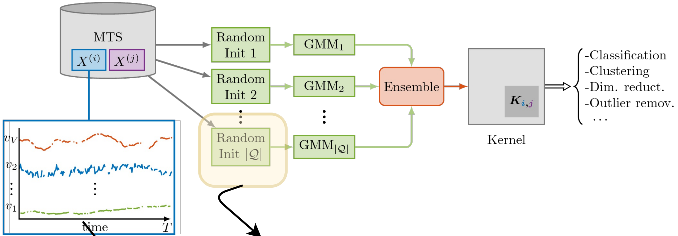




Randomly select

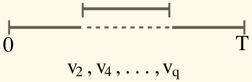
- Hyperparameters a_0, b_0, N_0
- Initialization of parameters μ, σ, θ
- Number of mixture components
- Subset of time series $\{X^{(n)}\}$

How do we do ensemble learning?

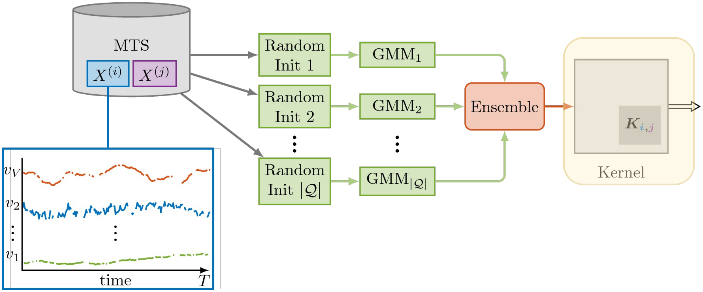


Randomly select

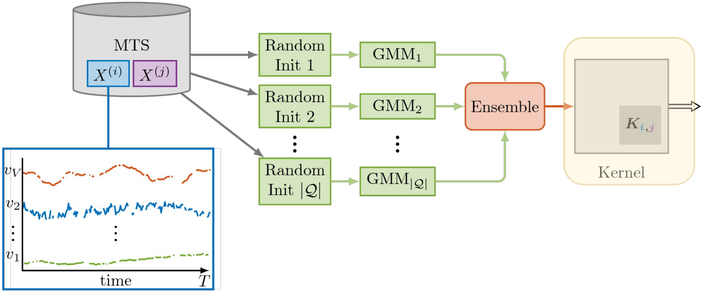
- Hyperparameters a_0, b_0, N_0
- Initialization of parameters μ, σ, θ
- Number of mixture components
- Subset of time series $\{X^{(n)}\}$
- Time segment, ,
- Subset of attributes



Forming the kernel



Forming the kernel



$$K(X^{(n)}, X^{(m)}) = \frac{1}{Z} \sum_{q \in \mathcal{Q}} \Pi^{(n)}(q)^T \Pi^{(m)}(q)$$

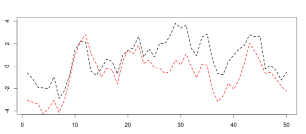
where
$$\Pi^{(n)}(q) \equiv (\pi_1^{(n)}, \ldots, \pi_{q_2}^{(n)})^T$$

and
$$\pi_g^{(n)} \equiv P(Z_g^{(n)} = 1 \mid X^{(n)},\ R^{(n)},\ \Theta)$$

Two-variate time series' generated from a VAR model

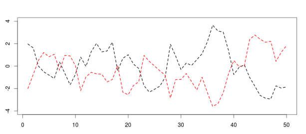
VAR(1) dataset

Class 1



Positive correlation.

Class 2

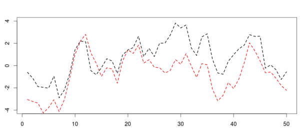


Negative correlation.

Two-variate time series' generated from a VAR model

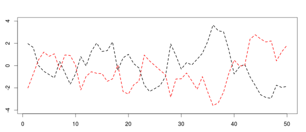
VAR(1) dataset

Class 1



Positive correlation.

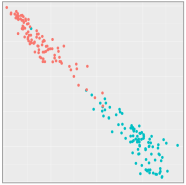
Class 2



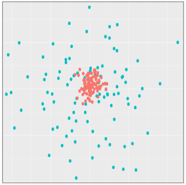
Negative correlation.

Dimensionality reduction using kPCA

TCK



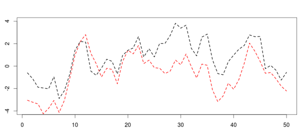
Lin. Kernel



Two-variate time series' generated from a VAR model

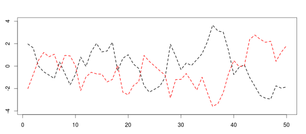
VAR(1) dataset

Class 1



Positive correlation.

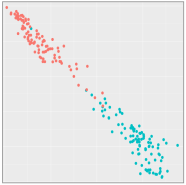
Class 2



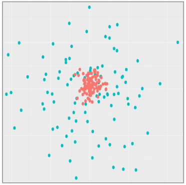
Negative correlation.

Dimensionality reduction using kPCA

TCK



Lin. Kernel



Clustering

	TCK	GMM
CA	0.990	0.910
ARI	0.961	0.671

Missing data

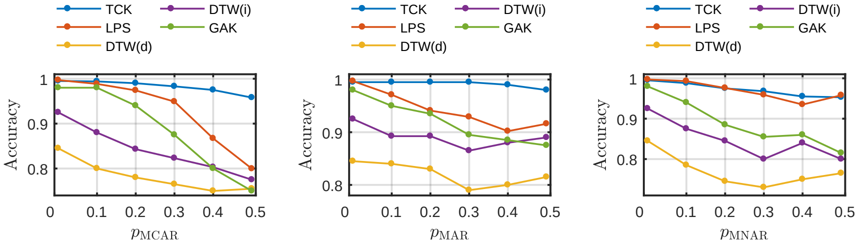


Figure: Classification accuracy on simulated VAR(1) dataset of the 1NN-classifier configured with a (dis)similarity matrix obtained using LPS, DTW (d), DTW (i), GAK and TCK. We report results for three different types of missingness, with an increasing percentage of missing values.

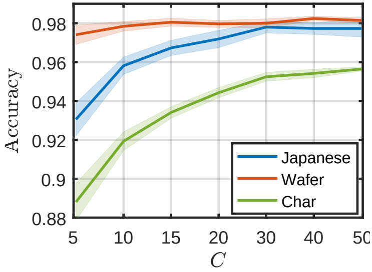
Benchmark datasets

Datasets	Attributes	Classes	Source	TCK	LPS	DTW
PenDigits	2	10	UCI	**0.929**	0.928	0.883
Libras	2	15	UCI	0.811	**0.894**	0.878
uWave	3	8	UCR	0.908	**0.945**	0.909
Character Trajectories	3	20	UCI	0.953	**0.961**	0.903
Robot failure LP1	6	4	UCI	**0.938**	0.836	0.720
Robot failure LP4	6	3	UCI	**0.926**	0.914	0.880
Wafer	6	2	UCR	**0.982**	0.981	0.963
Japanese vowels	12	9	UCI	**0.978**	0.964	0.965
ArabicDigits	13	10	UCI	0.951	**0.977**	0.962
PEMS	963	7	UCI	**0.815**	0.798	0.775
ItalyPower	1	2	UCR	**0.947**	0.933	0.918
Synthetic control	1	6	UCR	**0.993**	0.975	0.937

Table: Accuracy.

Robustness to hyperparameters

One (two) hyperparameters: Number of GMMs in the ensemble $|\mathcal{Q}|$.
$|\mathcal{Q}| = Q(C-1)$.



Outline

1. Introduction

2. Time series cluster kernel for learning similarities between multivariate time series with missing data

3. Time series cluster kernels to exploit informative missingness and incomplete label information

Paper III

K. Ø. Mikalsen, C. Soguero-Ruiz, F. M. Bianchi, A. Revhaug and R. Jenssen,
"**Time series cluster kernels to exploit informative missingness and incomplete label information**",
submitted to *Pattern Recognition*.

Informative missingness

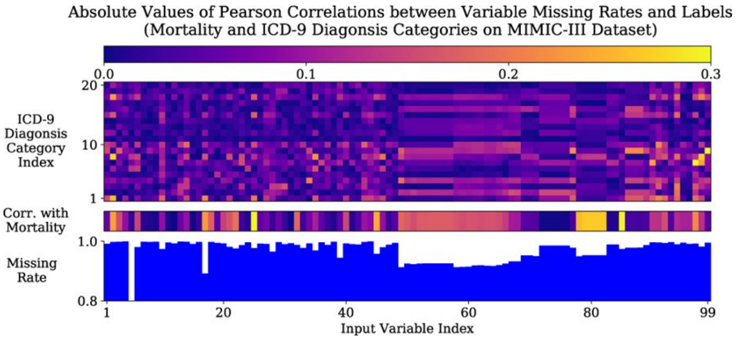
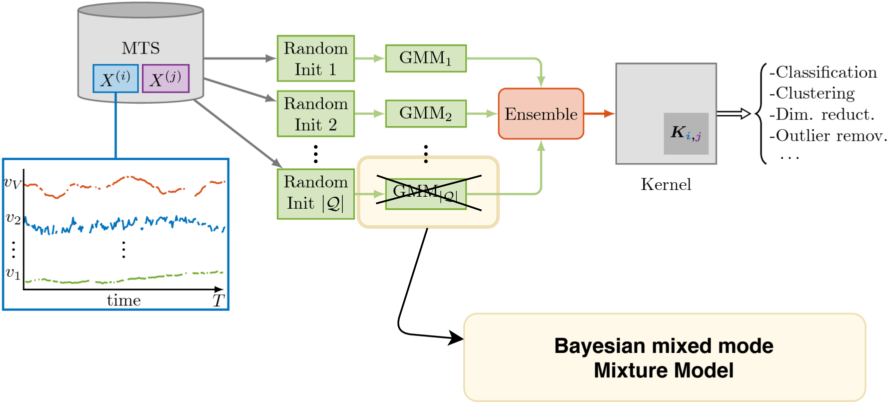


Figure: Che et al, 2018.

The distribution of the missing patterns for diseased patients is not equal to the corresponding distribution for the control group: $p(R \mid Y = 1) \neq p(R \mid Y = 0)$.

- Missingness not ignorable!
- TCK_{IM}: Exploit the rich information in the missingness patterns and observed data.

- Binary indicator time series.
- Continuous and discrete attributes.
- Mixed mode Bayesian mixture models.



Bayesian mixed mode mixture model

$X \in \mathbb{R}^{V \times T}$, input time series: $X = \begin{pmatrix} 13.1 & NA & NA & 14.2 & NA & 14.4 & NA \\ NA & 51 & 52 & NA & 40 & NA & 37 \end{pmatrix}$

$R \in \{0,1\}^{V \times T}$, masking for X: $R = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 & 1 \end{pmatrix}$

Bayesian mixed mode mixture model

$X \in \mathbb{R}^{V \times T}$, input time series: $X = \begin{pmatrix} 13.1 & NA & NA & 14.2 & NA & 14.4 & NA \\ NA & 51 & 52 & NA & 40 & NA & 37 \end{pmatrix}$

$R \in \{0,1\}^{V \times T}$, masking for X: $R = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 & 1 \end{pmatrix}$

Multivariate time series $U = (X, R)$ with two modes X and R.
U is generated from a finite mixture density,

$$p(U \mid \Phi, \Theta) = \sum_{g=1}^{G} \theta_g f(U \mid \phi_g),$$

Bayesian mixed mode mixture model

$X \in \mathbb{R}^{V \times T}$, input time series: $X = \begin{pmatrix} 13.1 & NA & NA & 14.2 & NA & 14.4 & NA \\ NA & 51 & 52 & NA & 40 & NA & 37 \end{pmatrix}$

$R \in \{0,1\}^{V \times T}$, masking for X: $R = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 & 1 \end{pmatrix}$

Multivariate time series $U = (X, R)$ with two modes X and R.
U is generated from a finite mixture density,

$$p(U \mid \Phi, \Theta) = \sum_{g=1}^{G} \theta_g f(U \mid \phi_g),$$

Assume that

$$f(U \mid \phi_g) = \underbrace{f(X \mid R, \mu_g, \Sigma_g)}_{Gaussian} \underbrace{f(R \mid \beta_g)}_{Bernoulli},$$

Bayesian mixed mode mixture model

$X \in \mathbb{R}^{V \times T}$, input time series: $X = \begin{pmatrix} 13.1 & NA & NA & 14.2 & NA & 14.4 & NA \\ NA & 51 & 52 & NA & 40 & NA & 37 \end{pmatrix}$

$R \in \{0,1\}^{V \times T}$, masking for X: $R = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 & 1 \end{pmatrix}$

Multivariate time series $U = (X, R)$ with two modes X and R.
U is generated from a finite mixture density,

$$p(U \mid \Phi, \Theta) = \sum_{g=1}^{G} \theta_g f(U \mid \phi_g),$$

Assume that

$$f(U \mid \phi_g) = \underbrace{f(X \mid R, \mu_g, \Sigma_g)}_{Gaussian} \underbrace{f(R \mid \beta_g)}_{Bernoulli},$$

Hence

$$f(X \mid R, \mu_g, \Sigma_g) = \prod_{v=1}^{V} \prod_{t=1}^{T} \mathcal{N}(x_v(t) \mid \mu_{gv}(t), \sigma_{gv})^{r_v(t)},$$

and

$$f(R \mid \beta_g) = \prod^{V} \prod^{T} \beta_{gvt}^{r_v(t)} (1 - \beta_{gvt})^{1 - r_v(t)}.$$

Algorithm 2 MAP-EM for mixed mode mixture model

Input Dataset $\{U^{(n)} = (X^{(n)}, R^{(n)})\}_{n=1}^{N}$, hyperparameters Ω and number of mixtures G.

1: Initialize the parameters $\Theta = (\theta_1, \dots, \theta_G)$ and $\Phi = \{\mu_g, \sigma_g, \beta_g\}_{g=1}^{G}$.

2: E-step. For each MTS $U^{(n)}$, evaluate the posterior probabilities with the current parameter estimates.

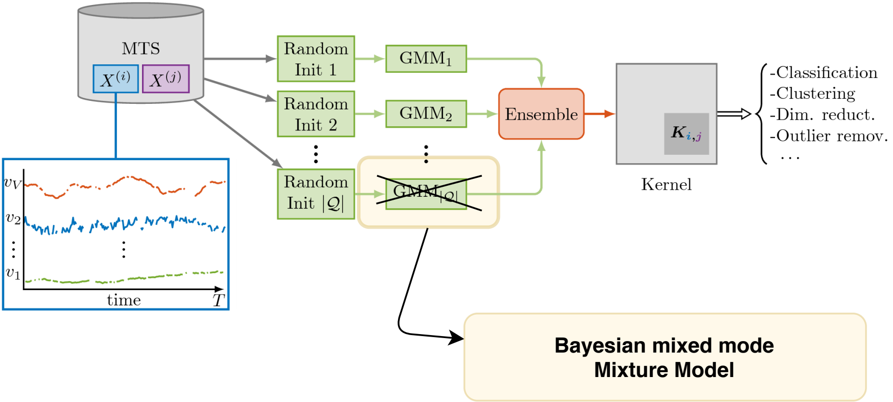
3: M-step. Update parameters using the current posteriors

$$
\begin{aligned}
\theta_g &= N^{-1} \textstyle\sum_{n=1}^{N} \pi_g^{(n)} \\
\sigma_{gv}^2 &= \frac{N_0 s_v^2 + \sum_{n=1}^{N} \sum_{t=1}^{T} r_v^{(n)}(t)\, \pi_g^{(n)} \left(x_v^{(n)}(t) - \mu_{gv}(t)\right)^2}{N_0 + \sum_{n=1}^{N} \sum_{t=1}^{T} r_v^{(n)}(t)\, \pi_g^{(n)}} \\
\mu_{gv} &= \frac{S_v^{-1} m_v + \sigma_{gv}^{-2} \sum_{n=1}^{N} \pi_g^{(n)} \mathrm{diag}(r_v^{(n)})\, x_v^{(n)}}{S_v^{-1} + \sigma_{gv}^{-2} \sum_{n=1}^{N} \pi_g^{(n)} \mathrm{diag}(r_v^{(n)})} \\
\beta_{gvt} &= (\textstyle\sum_{n=1}^{N} \pi_g^{(n)})^{-1} \sum_{n=1}^{N} \pi_g^{(n)} r_v^{(n)}(t)
\end{aligned}
$$

4: Repeat step 2-3 until convergence.

Output Posteriors $\Pi^{(n)} \equiv \left(\pi_1^{(n)}, \dots, \pi_G^{(n)}\right)^T$ and parameter estimates Θ and Φ.

TCK$_{IM}$



Create the kernel in the same way as in Paper 1, but with Bayesian mixed mode mixture models as base models in the ensemble!

Healthcare case study

Tromsø EHR corpus:

- Data extracted from Department of Gastrointestinal Surgery at University Hospital of North-Norway.
- More than 35000 unique patients and approximately 264 000 outpatient visits.
- Procedure codes: More than 1 000 000 NCSP codes.
- Diagnosis codes: More than 1 000 000 ICD-10 codes.
- Laboratory tests: More than 1 600 000 lab tests.
- Free text notes: More than 1 800 000. Hundreds of different document categories.
- Radiologic examinations: More than 60 000 radiology reports.
- Histology data: more than 500 000 pathology reports, including (re)-admittance and death dates.

Detecting infections among patients undergoing colon rectal cancer surgery

- Detect Surgical Site Infection (SSI), a common hospital-acquired infection.
- Laboratory tests $\rightarrow$ Multivariate time series.

Attribute nr.	Blood test	Missing rate
1	Hemoglobin	0.646
2	Leukocytes	0.727
3	C-Reactive Protein	0.691
4	Potassium	0.709
5	Sodium	0.712
6	Creatinine	0.867
7	Thrombocytes	0.921
8	Albumin	0.790
9	Carbamide	0.940
10	Glucose	0.921
11	Amylase	0.952

Overall: 80.7% missing data.

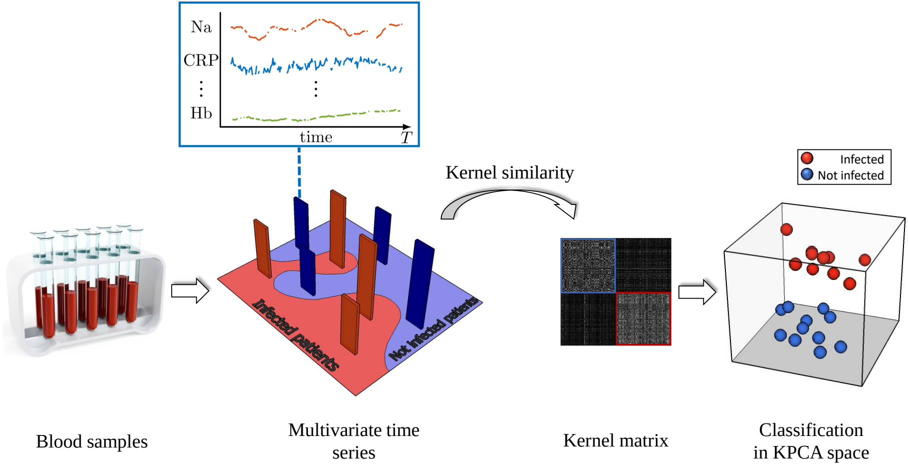


Figure: *Overview of the approach taken to detect postoperative SSI.*

Results

Table: Performance (mean $\pm$ se) on the SSI dataset.

Kernel	F1-score	Sensitivity	Specificity	Accuracy
TCK	0.726 ± 0.045	0.678 ± 0.035	0.930 ± 0.024	0.863 ± 0.023
LPS	0.746 ± 0.035	0.696 ± 0.056	0.939 ± 0.019	0.875 ± 0.016
GAK_{LOCF}	0.570 ± 0.045	0.484 ± 0.059	0.924 ± 0.022	0.808 ± 0.017
GAK_{mean}	0.629 ± 0.046	0.502 ± 0.059	$\mathbf{0.966} \pm \mathbf{0.023}$	0.843 ± 0.016
Linear_{LOCF}	0.557 ± 0.058	0.480 ± 0.073	0.914 ± 0.017	0.800 ± 0.018
Linear_{mean}	0.599 ± 0.030	0.489 ± 0.041	0.948 ± 0.043	0.826 ± 0.024
LPS_{IM}	0.720 ± 0.062	0.661 ± 0.069	0.937 ± 0.036	0.863 ± 0.032
$\text{GAK}_{IM+LOCF}$	0.669 ± 0.015	0.586 ± 0.024	0.940 ± 0.021	0.846 ± 0.011
$\text{GAK}_{IM+mean}$	0.696 ± 0.030	0.617 ± 0.033	0.945 ± 0.022	0.856 ± 0.011
$\text{Linear}_{IM+LOCF}$	0.628 ± 0.016	0.529 ± 0.030	0.945 ± 0.011	0.834 ± 0.005
$\text{Linear}_{IM+mean}$	0.668 ± 0.037	0.568 ± 0.033	0.951 ± 0.030	0.850 ± 0.021
TCK_{IM}	$\mathbf{0.802} \pm \mathbf{0.016}$	$\mathbf{0.806} \pm \mathbf{0.027}$	0.927 ± 0.017	$\mathbf{0.895} \pm \mathbf{0.010}$

Other MTS work

- K. Ø. Mikalsen, F. M. Bianchi, C. Soguero-Ruiz, R. Jenssen, "**The time series cluster kernel**", published in *MLSP 2017*, Tokyo, Japan, Sep. 2017, pp. 1–6.
- K. Ø. Mikalsen, F. M. Bianchi, C. Soguero-Ruiz, S. O. Skrøvseth, R.-O. Lindsetmo, A. Revhaug, R. Jenssen, " **Learning similarities between irregularly sampled short multivariate time series from EHRs**", oral presentation at *3rd ICPR Workshop on Pattern Recognition for Healthcare Analytics*, Cancun, Mexico, Dec. 2016.
- A. Storvik Strauman, F. M. Bianchi, K. Ø. Mikalsen, M. Kampffmeyer, C. Soguero-Ruiz, R. Jenssen, "**Classification of postoperative surgical site infections from blood measurements with missing data using recurrent neural networks**", published in *Proceedings of 2018 IEEE EMBS International Conference on Biomedical Health Informatics (BHI)*, Las Vegas, USA, Mar. 2018, pp 307–310.
- F. M. Bianchi, K. Ø. Mikalsen and R. Jenssen, "**Learning compressed representations of blood samples time series with missing data** ", ESANN, Bruges, Belgium, Apr. 2018,
- F. M. Bianchi, L. Livi, K. Ø. Mikalsen, M. Kampffmeyer and R. Jenssen, "**Learning representations of multivariate time series with missing data** ", *Pattern Recognition*, 2019.