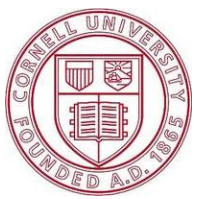
Online Model-Based Clustering for Crisis Identification in Distributed Computing

Dawn Woodard
School of Operations Research and Information Engineering
& Dept. of Statistical Science, Cornell University

with Moises Goldszmidt, Microsoft Research

Harvard University Statistics Department, 2011



Outline

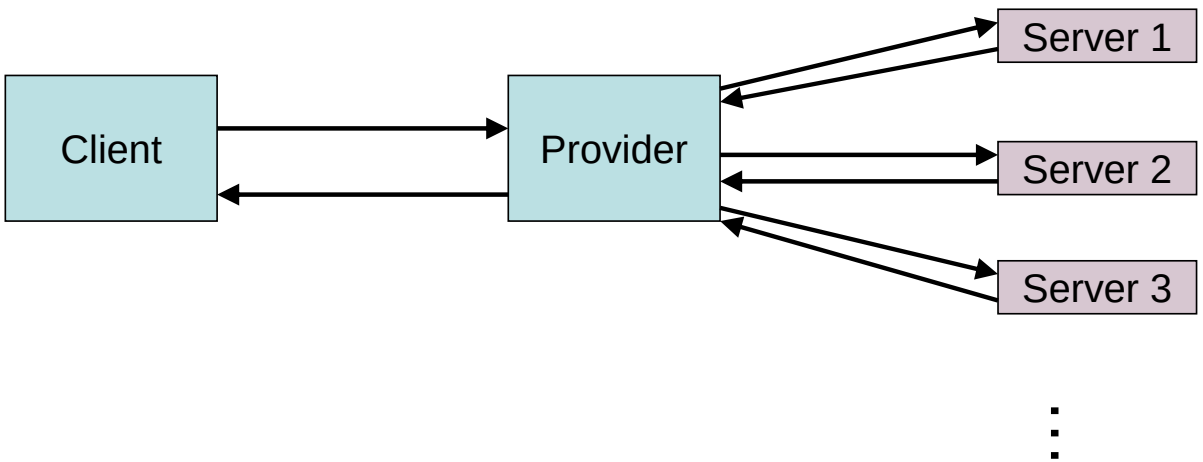
1. Background and Overview
2. Modeling
3. Computation and Decision Making
   - Offline Computation
   - Online Computation
   - Decision Making
4. Simulation Study
   - Offline
   - Online
5. Application to the Email Hosted Service
   - Offline
   - Online
6. Conclusions

Distributed Computing

- **Large distributed computing systems** provide the computing power behind internet services, cloud computing, and more; examples include search, email processing, e-commerce, and storage.
- Operate in datacenters hosting thousands to **tens of thousands of servers**
- E.g. Microsoft's Email Hosted Service (EHS)
  - 24/7 email processing incl. spam filtering, encryption

Distributed Computing

This processing is performed in parallel:



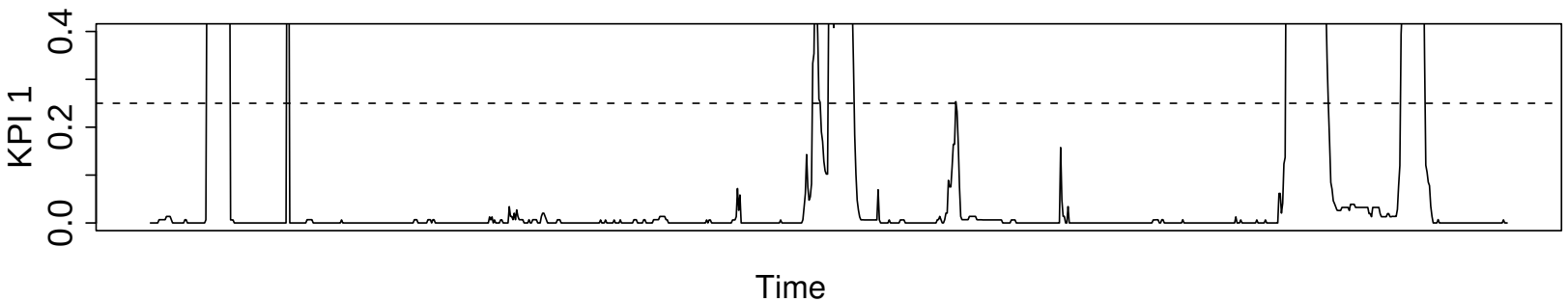
Distributed Computing

Distributed Computing

- Availability & responsiveness goals are inevitably compromised by hardware and software problems
- Can have occasional severe violation of performance goals (“crises”)
- E.g. due to:
  - servers becoming overloaded in periods of high demand
  - performance problems in lower-level computing centers on which the servers rely (e.g. for performing authentication)
- If the problem lasts for more than a few minutes, must pay cash penalties to clients, have potential loss of contracts

Distributed Computing

Fraction of servers violating a performance goal, for a 10-day period in EHS:



Exceeding the dotted line (contractually defined) constitutes a crisis.

Distributed Computing

Need to rapidly recognize the recurrence of a problem

- If an effective intervention is known for this problem, can apply it

Due to large scale and interdependence, manual problem diagnosis is difficult and slow

Have a set of status measurements for each server. E.g., for EHS:

- CPU utilization
- Memory utilization
- For each spam filter, the length of the queue and the throughput
- ...

Distributed Computing

- Goal: Match a currently occurring (i.e., incompletely observed) crisis to previous crises of mixed known and unknown causes
  - any previous crises have same type as the new crisis? Which ones?
- This is an online clustering problem with:
  - partial labeling
  - incomplete data for the new crisis
- We use model-based clustering based on a Dirichlet process mixture (e.g. Escobar & West 1995)
  - allows estimation of # of clusters
- The evolution of each crisis is modeled as a multivariate time series

Cost-Optimal Decision Making

Wish to perform optimal (expected-cost-minimizing) decision making during a crisis...

...while accounting for uncertainty in the crisis type assignments and the parameters of those types

This requires fully Bayesian inference

Fully Bayesian Inference

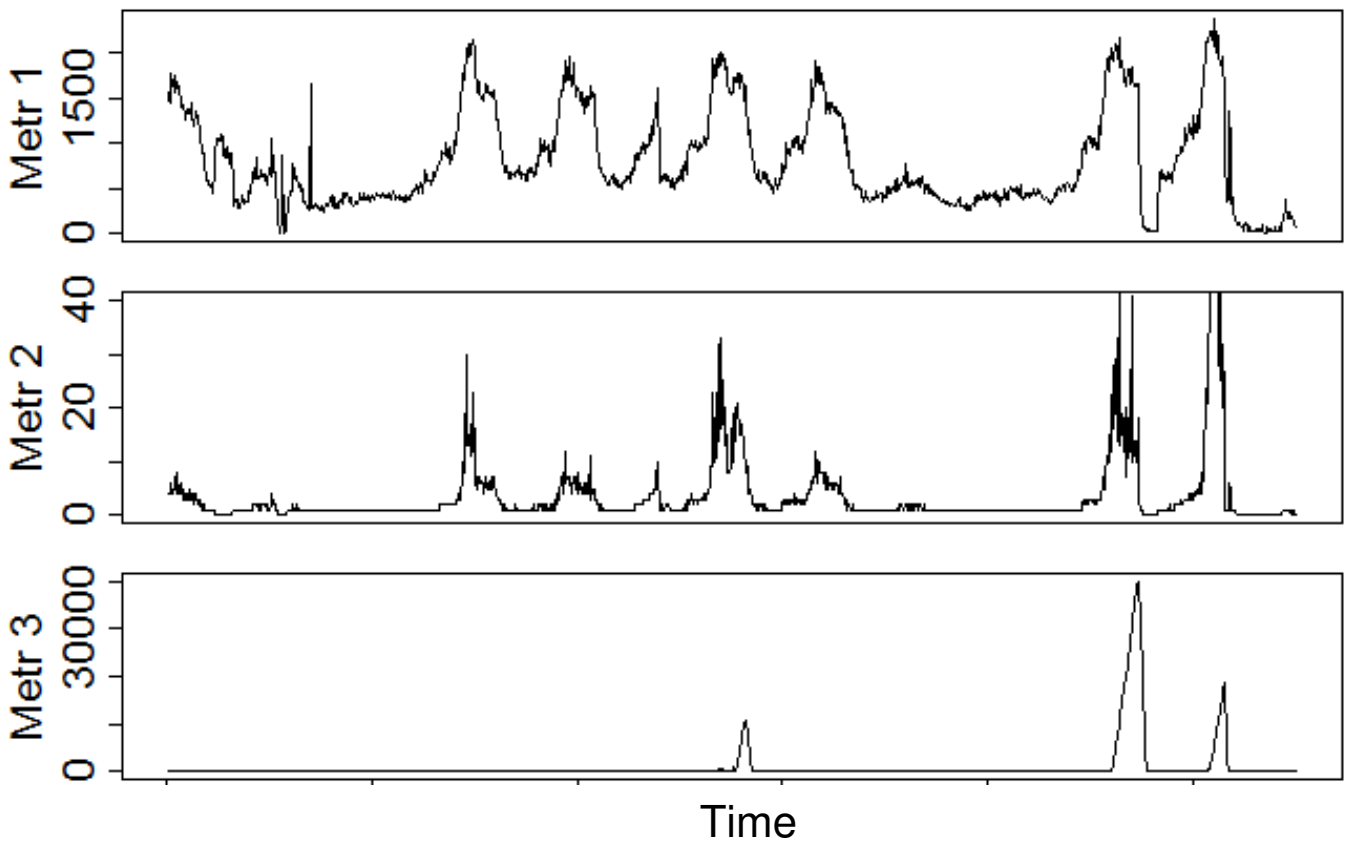
- We apply fully Bayesian inference (via MCMC) in the periods between crises
  - Due to posterior multimodality, we combine a collapsed-space split-merge method with parallel tempering
- As a new crisis begins, do fast Bayesian prediction

Related Work

- Ours is the first instance of fully Bayesian real-time online clustering without use of a variational approximation
  - Unlike VB we capture the multiple modes & dependencies in the posterior dist'n
- Online model-based clustering of documents / images: Sato (2001); Zhang, Ghahramani, & Yang (2004); Gomez, Welling, & Perona (2008)
  - variational approximation to posterior dist'n
- Fully Bayesian clustering: Bensmail et al. (1997); Pritchard, Stephens, & Donnelly (2000); Lau & Green (2007)
- Many examples of fully Bayesian mixture modeling

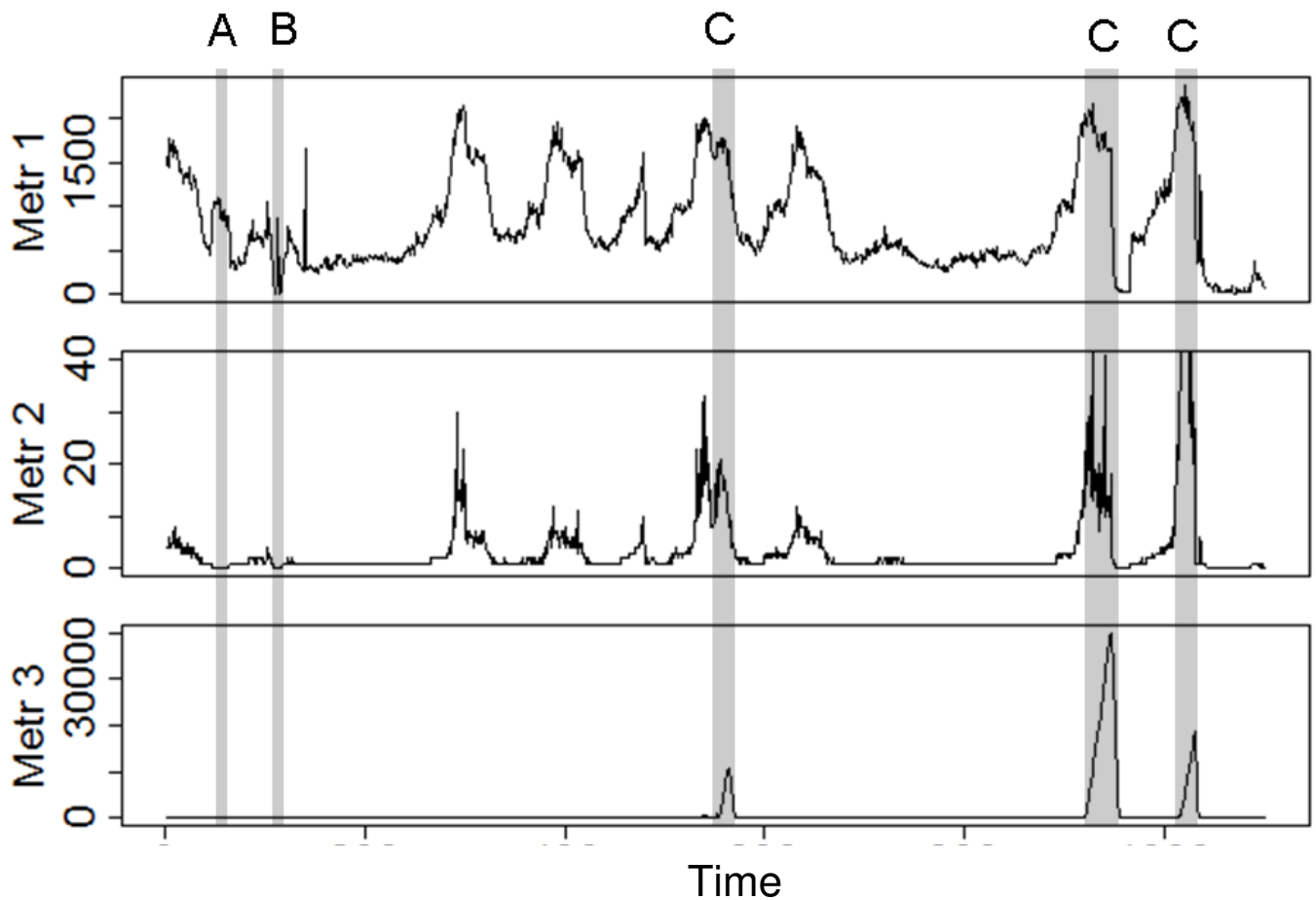
Data

Medians of 3 metrics (e.g. CPU, memory util.) across servers, for a 10-day period (EHS):



Data

Crises are highlighted; letters indicate their known type:



Data

- The medians of the metrics are very informative as to crisis type
- Specifically, whether the median is low, normal, or high
- We fit our models to the median values of the metrics, discretized into 1: low, 2: normal, and 3: high

Crisis Modeling

Time series model for crisis evolution:

- $Y_{ij\ell}$: value of metric j in the ℓth time period after the start of crisis i
- Assume metrics independent conditional on crisis type (for parsimony)
- For crisis type k, Y_{ij1} is drawn from a discrete dist'n with probability vector $\gamma^{(jk)}$
- ...and $Y_{ij\ell}$ evolves according to a Markov chain with transition matrix $T_{\cdot\cdot}^{(jk)}$

Crisis Modeling

$\Rightarrow$ Complete-data likelihood fn:

$$\pi\left(\mathcal{D} \mid \{Z_i\}_{i=1}^{I}, \{\gamma^{(jk)}, T_{\cdot\cdot}^{(jk)}\}_{j,k}\right) = \prod_{i,j,t}\left[\left(\gamma_t^{(j\,Z_i)}\right)^{\mathbf{1}(Y_{ij1}=t)} \prod_{s}\left(T_{st}^{(j\,Z_i)}\right)^{n_{ijst}}\right].$$

conditioning on the unknown type indicators Z_i of each crisis $i = 1, \ldots, I$.

n_{ijst}: the number of transitions of the jth metric from state s to state t during crisis i

Cluster Modeling

Dirichlet process mixture (DPM) model:

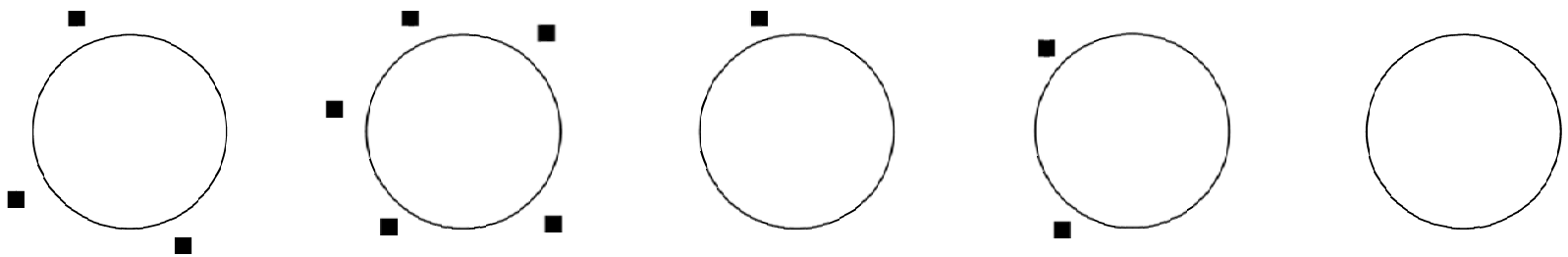
- Natural for online clustering
  - Allows estimation of # of clusters
  - Observations are exchangeable
- Parameterized by
  - α: controls the expected number of clusters occurring in a fixed number of observations
  - G_0: the prior $G_0(\{\gamma^{(j\cdot)}, T_{\cdot\cdot}^{(j\cdot)}\}_j)$ for the parameters associated with each cluster k

Cluster Modeling

Also called the "Chinese Restaurant Process":

$$\pi\left(Z_i = k \mid \{Z_{i'}\}_{i'<i}\right) \propto \begin{cases} \alpha & : \quad k \text{ is a new type} \\ \#\{i' < i : Z_{i'} = k\} & : \quad \text{else} \end{cases}$$

Each observation i is a new guest who either sits at an occupied table with prob. proportional to the number of guests at that table, or sits at an empty table:



Guests at same table share same dishes, i.e. have same parameters.

Cluster Modeling

Conditional on $\{Z_i\}_{i=1}^{I}$, parameters of the clusters are independently dist'ed according to G_0:

$$\pi\left(\{\gamma^{(jk)}, T_{\cdot\cdot}^{(jk)}\}_{j,k} \mid \{Z_i\}_{i=1}^{I}\right) = \prod_{k=1}^{m_I} G_0\left(\{\gamma^{(jk)}, T_{\cdot\cdot}^{(jk)}\}_j\right).$$

Cluster Modeling

Now we have an expression for the posterior density of all unknowns:

$$\pi\left(\{Z_i\}_{i=1}^{I}, \{\gamma^{(jk)}, T_{\cdot\cdot}^{(jk)}\}_{j,k} \mid \mathcal{D}\right) \propto$$

$$\pi\left(\{Z_i\}_{i=1}^{I}\right) \pi\left(\{\gamma^{(jk)}, T_{\cdot\cdot}^{(jk)}\}_{j,k} \mid \{Z_i\}_{i=1}^{I}\right) \pi\left(\mathcal{D} \mid \{Z_i\}_{i=1}^{I}, \{\gamma^{(jk)}, T_{\cdot\cdot}^{(jk)}\}_{j,k}\right)$$

Cluster Modeling

Partially labeled case:

- Can capture partial labelling info. with indicators $\mathbf{1}(Z_i = Z_{i'})$ for some pairs $i \sim i'$ and $\mathbf{1}(Z_i \neq Z_{i'})$ for other pairs $i \not\sim i'$
- Multiply prior by $\prod_{i \sim i'} \mathbf{1}(Z_i = Z_{i'}) \prod_{i \not\sim i'} \mathbf{1}(Z_i \neq Z_{i'})$
- Our comp. method extends trivially, by disallowing configurations that are incompatible with the partial labelling.

Cluster Modeling

G_0:

- Independent Dirichlet priors for $\gamma^{(jk)}$
- Independent product Dirichlet priors for $T_{\cdot\cdot}^{(jk)}$

Offline Computation

- The cluster parameters $\{\gamma^{(jk)}, T_{\cdot\cdot}^{(jk)}\}_{j,k}$ can be integrated analytically out of the posterior
- Run a Markov chain with target dist'n $\pi(\{Z_i\}_{i=1}^{I} \mid \mathcal{D})$
- Jain and Neal (2004) use a Gibbs sampler, with an additional split-merge move on clusters
- We add parallel tempering (Geyer 1991)

Online Inference

- Wish to identify a crisis in real time
- Have data $\mathcal{D}$ from previous crises and data $\mathcal{D}_{new}$ so far for the new crisis
- E.g., wish to estimate $\pi(Z_{new} = Z_i \mid \mathcal{D}, \mathcal{D}_{new})$ for each previous crisis $i = 1, \ldots, I$
- ...and $\pi(Z_{new} \neq Z_i \ \forall i \mid \mathcal{D}, \mathcal{D}_{new})$

Exact Online Inference

Method 1:

- Just apply the Markov chain method to the data from the $I + 1$ crises
- Gives posterior sample vectors $\left(\{Z_i^{(\ell)}\}_{i=1}^{I}, Z_{new}^{(\ell)}\right)$ for $\ell = 1, \ldots, L$
- Monte Carlo estimates of the desired probabilities:

$$\hat{\pi}(Z_{new} = Z_i \mid \mathcal{D}, \mathcal{D}_{new}) = \frac{1}{L} \sum_{\ell=1}^{L} \mathbf{1}(Z_{new}^{(\ell)} = Z_i^{(\ell)})$$

$$\hat{\pi}(Z_{new} \neq Z_i \ \forall i \mid \mathcal{D}, \mathcal{D}_{new}) = \frac{1}{L} \sum_{\ell=1}^{L} \mathbf{1}(Z_{new}^{(\ell)} \neq Z_i^{(\ell)} \ \forall i)$$

- But running the Markov chain is too slow for real-time decision making!

Fast Online Prediction

Method 2:

We give a method using the predictive approximation:

$$\pi(Z_{new} = Z_i \mid \mathcal{D}, \mathcal{D}_{new}) = \sum_{\{Z_{i'}\}_{i'=1}^{I}} \pi(Z_{new} = Z_i \mid \{Z_{i'}\}_{i'=1}^{I}, \mathcal{D}, \mathcal{D}_{new}) \pi(\{Z_{i'}\}_{i'=1}^{I} \mid \mathcal{D}, \mathcal{D}_{new})$$

$$\approx \sum_{\{Z_{i'}\}_{i'=1}^{I}} \pi(Z_{new} = Z_i \mid \{Z_{i'}\}_{i'=1}^{I}, \mathcal{D}, \mathcal{D}_{new}) \pi(\{Z_{i'}\}_{i'=1}^{I} \mid \mathcal{D})$$

* Assumes that $\mathcal{D}_{new}$ does not tell us much about the past crisis types

Fast Online Prediction

Method 2: Fast Online Inference

1. After the end of each crisis, rerun the Markov chain, yielding sample vectors $\{Z_i^{(\ell)}\}_{i=1}^{I}$ from the posterior $\pi(\{Z_i\}_{i=1}^{I} \mid \mathcal{D})$.

2. When a new crisis begins, use its data $\mathcal{D}_{new}$ to calculate the Monte Carlo estimates:

$$\hat{\pi}(Z_{new} = Z_i \mid \mathcal{D}, \mathcal{D}_{new}) = \frac{1}{L}\sum_{\ell=1}^{L} \pi(Z_{new} = Z_i^{(\ell)} \mid \{Z_{i'}^{(\ell)}\}_{i'=1}^{I}, \mathcal{D}, \mathcal{D}_{new})$$

$$\hat{\pi}(Z_{new} \neq Z_i \ \forall i \mid \mathcal{D}, \mathcal{D}_{new}) = \frac{1}{L}\sum_{\ell=1}^{L} \pi(Z_{new} \neq Z_i^{(\ell)} \ \forall i \mid \{Z_{i'}^{(\ell)}\}_{i'=1}^{I}, \mathcal{D}, \mathcal{D}_{new}).$$

(RHS available in closed form)

Fast Online Prediction

Part 2 is O(LIJ), very fast

Optimal Decision Making

- Want expected-cost-minimizing decision making during a crisis
- The total cost of the new crisis is a function $C(\phi, Z^*_{new})$ of:
  - The intervention ϕ
  - The true type Z^*_{new} of the current crisis
- Finding the expected cost of the crisis for intervention ϕ requires integrating C over the posterior distribution of Z_{new}
- Can be done exactly using Method 1, or approximately using Method 2

Simulation Study

Offline:

- Simulate I crises from a finite mixture model; apply our method (DPM) to all crises together
- **Compare with maximimum likelihood inference in a finite mixture model** ("ML-BIC"; Fraley & Raftery 2002):
  - Expectation-maximization to get MLE
  - Bayesian Information Criterion to choose # clusters
  - Initial clustering from hierarchical agglomerative clustering
- Also tried distance-based clustering, which did terribly

Simulation Study

Offline Accuracy Criteria:

1. **Pairwise Sensitivity:** For pairs of crises of the same type, % having prob. > 0.5 of being in the same cluster.
2. **Pairwise Specificity:** For pairs of crises not of the same type, % having prob. ≤ 0.5 of being in the same cluster.
3. **Error of No. Crisis Types:** The % error of the estimated number of crisis types
   - for DPM, post. mean is used to estimate # of types.

Simulation Study

No. Crises	No. Metrics	Method	Pairwise Sensitivity	Pairwise Specificity	% Error No. Types
15	**10**	**DPM**	**96.6 (1.45)**	**99.5 (0.29)**	**5.3 (1.22)**
		ML-BIC	54.0 (5.21)	98.0 (0.54)	77.4 (27.96)
15	**15**	**DPM**	**98.5 (0.90)**	**99.9 (0.05)**	**8.9 (3.71)**
		ML-BIC	39.8 (4.81)	99.9 (0.10)	113.0 (32.97)
25	**10**	**DPM**	**94.6 (2.49)**	**99.8 (0.10)**	**7.6 (1.62)**
		ML-BIC	59.1 (4.78)	98.6 (0.31)	24.2 (6.11)
25	**15**	**DPM**	**99.7 (0.32)**	**99.7 (0.19)**	**2.7 (0.84)**
		ML-BIC	40.9 (4.11)	99.8 (0.07)	86.0 (15.0)
35	**10**	**DPM**	**93.1 (1.43)**	**99.6 (0.09)**	**8.2 (1.68)**
		ML-BIC	61.2 (4.04)	98.0 (0.24)	35.0 (9.81)
35	**15**	**DPM**	**97.9 (0.95)**	**99.9 (0.06)**	**3.0 (0.60)**
		ML-BIC	46.2 (3.56)	99.7 (0.09)	51.8 (9.81)

Simulation Study

- DPM does far better than ML-BIC
  - ML-BIC cluster assignments rarely change much from their initial values
  - EM stuck in local modes
- More metrics $\Rightarrow$ better accuracy of DPM & worse accuracy of ML-BIC
- Tried several changes to ML-BIC, with little improvement:
  - smooth the initialization
  - smooth surface over which maximizing, by using a prior and getting MAP estimate instead of MLE

Simulation Study

Online:

- Compare Method 1 (“DPM-EX”) to Method 2 (“DPM”)

Simulation Study

Online Accuracy Criteria:

1. **Full-data misclassification rate:** % of crises with incorrect predicted type, using all of the data for the new crisis.
2. **p-period misclassification rate:** % of crises with incorrect predicted type, using the first p time periods of data for the new crisis.
3. **Average time to correct identification:** Avg. No. of time periods required to obtain the correct identification

("correct" predicted type: $\hat{\pi}(Z_{new} \neq Z_i \ \forall i \mid \mathcal{D}, \mathcal{D}_{new}) > 0.5$ if $Z^*_{new} \neq Z^*_i \ \forall i$ and otherwise $\hat{\pi}(Z_{new} = Z_i \mid \mathcal{D}, \mathcal{D}_{new}) > 0.5$ for some $i \leq I$ such that $Z^*_{new} = Z^*_i$)

Simulation Study

Online Accuracy:

No. Crises	No. Metrics	Method	Full-data Misclassification	3-period Misclassification	Avg. Time to Identification
15	**10**	**DPM**	**6.7 (3.0)**	**10.7 (4.5)**	**1.31 (0.11)**
		DPM-EX	8 (2.5)	10.7 (4.5)	–
15	**15**	**DPM**	**6.7 (5.2)**	**9.3 (6.2)**	**1.13 (0.08)**
		DPM-EX	5.3 (3.9)	8.0 (4.9)	–
25	**10**	**DPM**	**13.6 (2.7)**	**15.2 (2.7)**	**1.33 (0.13)**
		DPM-EX	9.6 (2.0)	15.2 (3.4)	–
25	**15**	**DPM**	**2.4 (1.6)**	**4.0 (1.8)**	**1.15 (0.06)**
		DPM-EX	3.2 (1.5)	3.2 (1.5)	–

Simulation Study

- Classification accuracy high ($> 80\%$) for both DPM & DPM-EX
- DPM not significantly worse than DPM-EX
- 3-period misclassification is not much $>$ than full-data misclassification
- Very early identification!

Application to EHS

27 crises in EHS during Jan-Apr 2008.

The causes of some of these were diagnosed later:

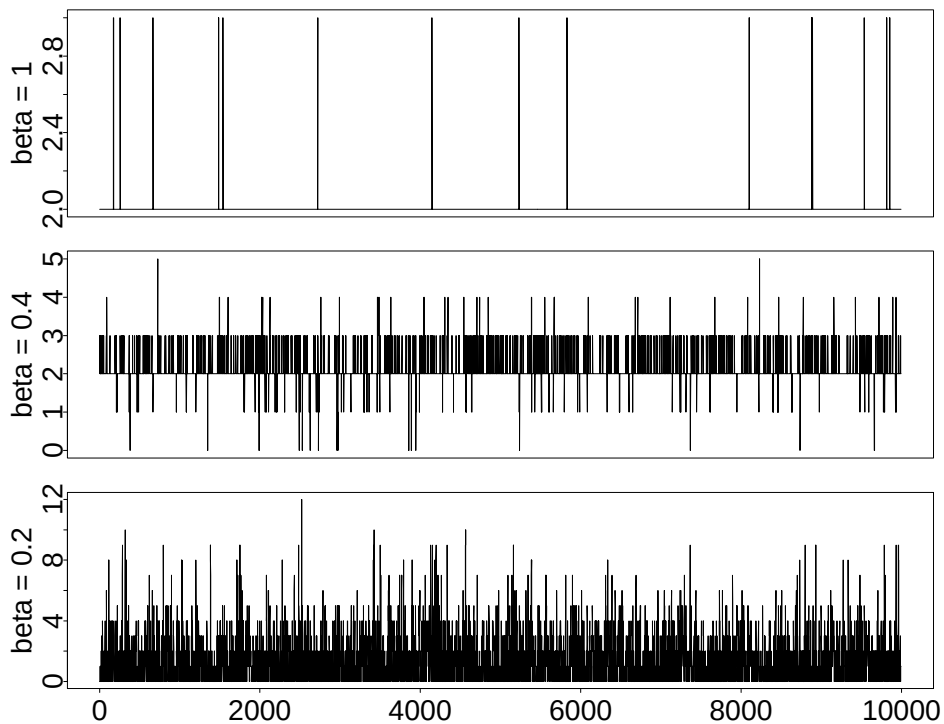
ID	Cause	No. of known crises
A	overloaded front-end	2
B	overloaded back-end	8
C	database configuration error	1
D	configuration error	1
E	performance issue	1
F	middle-tier issue	1
G	whole DC turned off and on	1
H	workload spike	1
I	request routing error	1

Offline Application to EHS

- Apply the Markov chain method to the set of 27 crises without the labels
- Compare to those labels

Offline Application to EHS

Trace plots of parallel tempering Markov chain samples of Z_{22}:



Geweke diag. p-value: 0.44

Gelman-Rubin scale factor: 1.01

Offline Application to EHS

Post. mode cluster assignment has 58% prob.

Sizes of clusters:

ID	Cause	No. of known crises	No. identified by DPM	No. DPM crises matching known
A	overloaded front-end	2	3	2
B	overloaded back-end	8	14	8
C	database configuration error	1	2	1
D	configuration error	1	0	0 (labeled as A)
E	performance issue	1	0	0 (labeled as B)
F	middle-tier issue	1	0	0 (labeled as I)
G	whole DC turned off and on	1	0	0 (labeled as B)
H	workload spike	1	1	1
I	request routing error	1	6	1

Offline Application to EHS

- Post. mode crisis labels mostly match known clusters
- The largest 5 clusters are correctly labelled
- Four uncommon crisis types are clustered with more common types
  - Crises having different causes can have the same patterns in their metrics
  - Need to add metrics that distinguish these types effectively

Online Application to EHS

Evaluate online accuracy, treating the posterior mode from the offline context as the gold standard.

- Original ordering:
  1. Full-data misclassification: 7.4%
  2. 3-period misclassification: 14.8%
  3. Avg. time to correct iden.: 1.81
- Permuting the crises:
  1. Full-data misclassification: 5.9% (SE =3.4%)
  2. 3-period misclassification: 11.8% (SE =3.2%)
  3. Avg. time to correct iden.: 1.56 (SE =0.07)

Conclusions

- Gave a method for fully Bayesian real-time crisis identification in distributed computing
- Described how to use this to perform rapid expected-cost-minimizing crisis intervention
- Very accurate on both simulated data and data from a production computing center

Reference: Woodard & Goldszmidt (2010). “Online model-based clustering for crisis identification in distributed computing.” JASA, In press.

References


Escobar, M. D. and West, M. (1995).
Bayesian density estimation and inference using mixtures.
Journal of the American Statistical Association, 90, 577-588.


Geyer, C. J. (1991).
Markov chain Monte Carlo maximum likelihood.
in *Computing Science and Statistics, Vol. 23: Proc. of the 23rd Symp. on the Interface*, ed. E. Keramidas, pp. 156-163.


Jain, S. and Neal, R. M. (2004).
A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model.
Journal of Computational and Graphical Statistics, 13, 158-182.


Lau, J. W. and Green, P. J. (2007).
Bayesian model-based clustering procedures.
Journal of Computational and Graphical Statistics, 16, 526-558.


Zhang, J., Ghahramani, Z., and Yang, Y. (2004).
A probabilistic model for online document clustering with application to novelty detection.
in *Advances in Neural Information Processing Systems*, ed. Y. Weiss.

Cluster Modeling

The DPM prior for the cluster indicators $\{Z_i\}_{i=1}^I$ and the cluster parameters $\gamma^{(jk)}$, $T_{\cdot\cdot}^{(jk)}$:

$$\pi\left(\{Z_i\}_{i=1}^I\right) = \prod_{i=1}^I \pi\left(Z_i \mid \{Z_{i'}\}_{i'<i}\right)$$

$$= \prod_{i=1}^I \left[\frac{\alpha}{\alpha+i-1}\mathbf{1}\left(Z_i=m_{i-1}+1\right)+\frac{1}{\alpha+i-1}\sum_{i'<i}\mathbf{1}\left(Z_i=Z_{i'}\right)\right]$$

where $m_i = \max\{Z_{i'} : i' \leq i\}$ for $i > 0$ and $m_0 = 0$.

Prior Constants

- Prior hyperparameters chosen by combining information in data with expert opinion
- Reflect the fact that the server status measurements are chosen to be indicative of crisis type
- Results far better than a “default” prior specification, which contradicts data and experts

Prior Constants

α:

- Prob. that 2 randomly chosen crises are of same type: $1/(\alpha+1)$
- EHS experts estimate as 0.1, giving $\alpha = 9$
- $\Rightarrow \sim$13 types in 27 crises

$\gamma^{(jk)} \sim \text{Dir}(a^{(j)})$. To choose $a^{(j)}$:

- Prior mean of $\gamma^{(jk)}$ taken as empirical dist'n of Y_{ij1} over i and j
- Substantial prob. that one of the $\gamma^{(jk)}$ is "close" to 1:

$$\pi\left((\gamma_1^{(jk)} > .85) \text{ OR } (\gamma_2^{(jk)} > .95) \text{ OR } (\gamma_3^{(jk)} > .85)\right) = 0.5$$

Analogous for $T_{\cdot\cdot}^{(jk)}$

Optimal Decision Making

- Want expected-cost-minimizing decision making during a crisis
- The total cost of the new crisis is a function $C\left[\phi, \{Z_i^*\}_{i=1}^I, Z_{new}^*\right]$ of:
  - The intervention ϕ
  - The true type Z_{new}^* of the current crisis
  - The vector of past crisis types $\{Z_i^*\}_{i=1}^I$, which give the context for Z_{new}^*

Optimal Decision Making

- If we knew C,
- given posterior sample vectors $\left(\{Z_i^{(l)}\}_{i=1}^{I}, Z_{new}^{(l)}\right)$ from the exact Method 1...
- ...the expected cost can be estimated as:

$$\mathbf{E}(C) \approx \frac{1}{L}\sum_{l=1}^{L} C\left[\phi, (\{Z_i^{(l)}\}_{i=1}^{I}, Z_{new}^{(l)})\right].$$

- Have a similar expression for approximate inferences from Method 2

Optimal Decision Making

- Don't know C in practice
- For interventions ϕ taken during previous crises can estimate C from realized costs
- Otherwise can estimate C from expert knowledge

Optimal Decision Making

- Since the goal is optimal intervention
- ...and since this requires the entire posterior distribution over $\left(\{Z_i\}_{i=1}^{I}, Z_{new}\right)$...
- we will avoid choosing a "best" cluster assignment
- instead focusing on the accuracy of the "soft identification", i.e. the posterior distribution over $\left(\{Z_i\}_{i=1}^{I}, Z_{new}\right)$

Simulation Study

K-means:

- Criteria for choosing the number of clusters do not work well in our context
- So we apply K-means using the true number of clusters (“K-means 1”)
- and half the true number of clusters (“K-means 2”)
- This is unrealistically optimistic...
- ...but K-means still does terribly