Concordia University
Engineering and Computer Science

A Bayesian Intrusion Detection Framework

Shuai FU & Dr. Nizar Bouguila

Presented by

Dr. Nizar Bouguila

For Cyber Security 2018 Scotland, UK

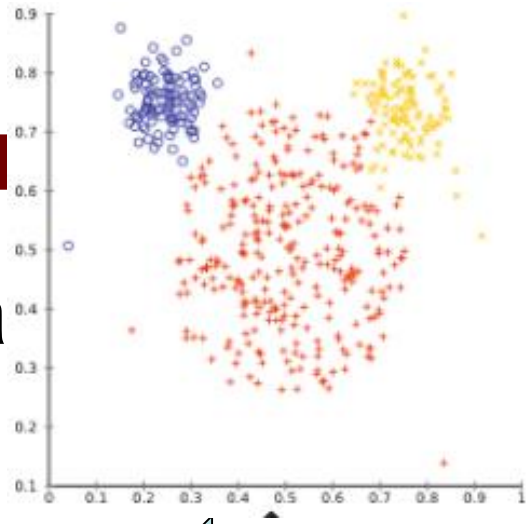
June 2018

f_shuai@encs.concordia.ca

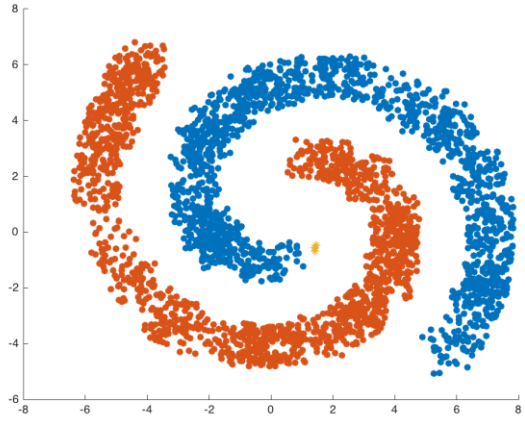
bouguila@ciise.concordia.ca

Université Concordia University

What's Machine Lear

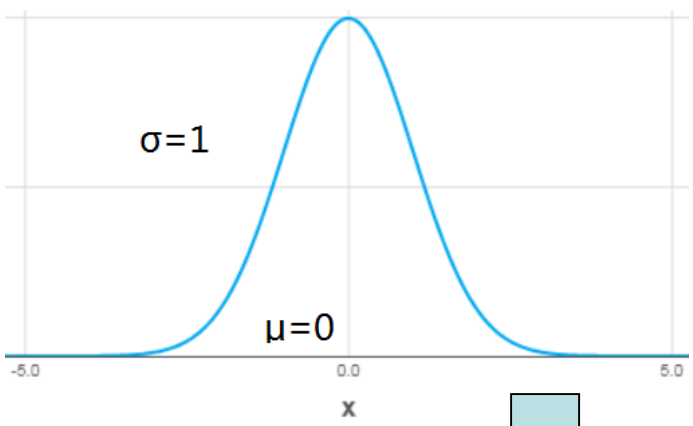


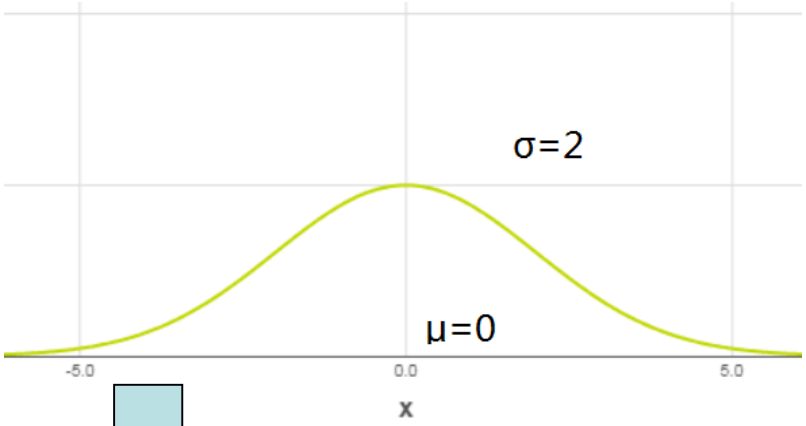
- "Learn" information from data
  - Pattern Recognition
    - Distance-based: K-means,…
    - Connectivity-based: DB-scan,…
    - Probability-based
    - …….

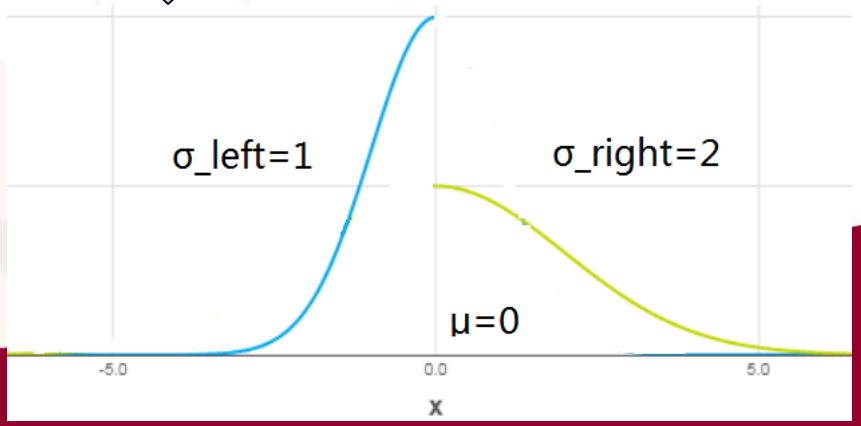


Asymmetric Gaussian

- Joint of two Gaussian distributions for every dimension







Asymmetric Gaussian Cont.

- Probability density function (PDF)
  - Dimension-by-dimension

$$p(X_n|\xi_j) \propto \prod_{k=1}^{d} \frac{1}{(\sigma_{l_{jk}} + \sigma_{r_{jk}})} \times \begin{cases} \exp\left[-\frac{(x_{nk}-\mu_{jk})^2}{2(\sigma_{l_{jk}})^2}\right] & if \ x_{nk} < \mu_{jk} \\ \exp\left[-\frac{(x_{nk}-\mu_{jk})^2}{2(\sigma_{r_{jk}})^2}\right] & if \ x_{nk} \geqslant \mu_{jk} \end{cases}$$

AGMM

- M-dimensional membership vector Z

$$Z_{ij} = \begin{cases} 1 & \text{if } X_i \text{ belongs to component } j \\ 0 & \text{otherwise} \end{cases}$$

- Weight parameter p_j $(0 < p_j \leq 1 \text{ and } \sum_{j=1}^{M} p_j = 1)$
- Likelihood function of AGMM

$$p(\mathcal{X}, Z|\Theta) = \prod_{i=1}^{N} \prod_{j=1}^{M} (p_j p(X_i|\xi_j))^{Z_{ij}}$$

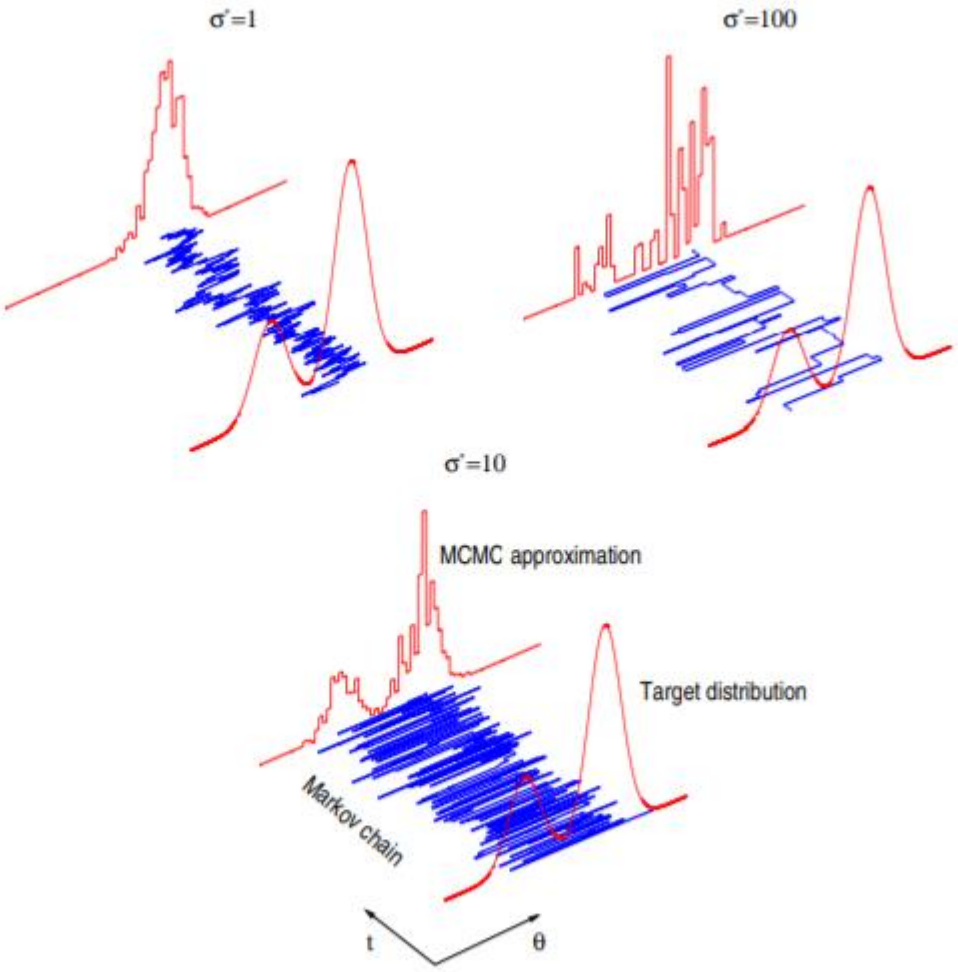
MCMC Implementations

- Metropolis-Hastings Algorithm
  - Sampling from proposal distribution
  - Sampling procedure is guided by an acceptance ratio to make a decision whether new status will be accepted or rejected

$$A_k(\mathbf{z}^\star, \mathbf{z}^{(\tau)}) = \min\left(1, \frac{\widetilde{p}(\mathbf{z}^\star) q_k(\mathbf{z}^{(\tau)}|\mathbf{z}^\star)}{\widetilde{p}(\mathbf{z}^{(\tau)}) q_k(\mathbf{z}^\star|\mathbf{z}^{(\tau)})}\right)$$

1.

2.



$$\frac{'|x)}{x')}$$

Figure 3: An example of the Metropolis Hastings algorithm for sampling from a mixture of two 1D Gaussians using a Gaussian proposal with different variances. Source: [AdFDJ03].

- (

d

1

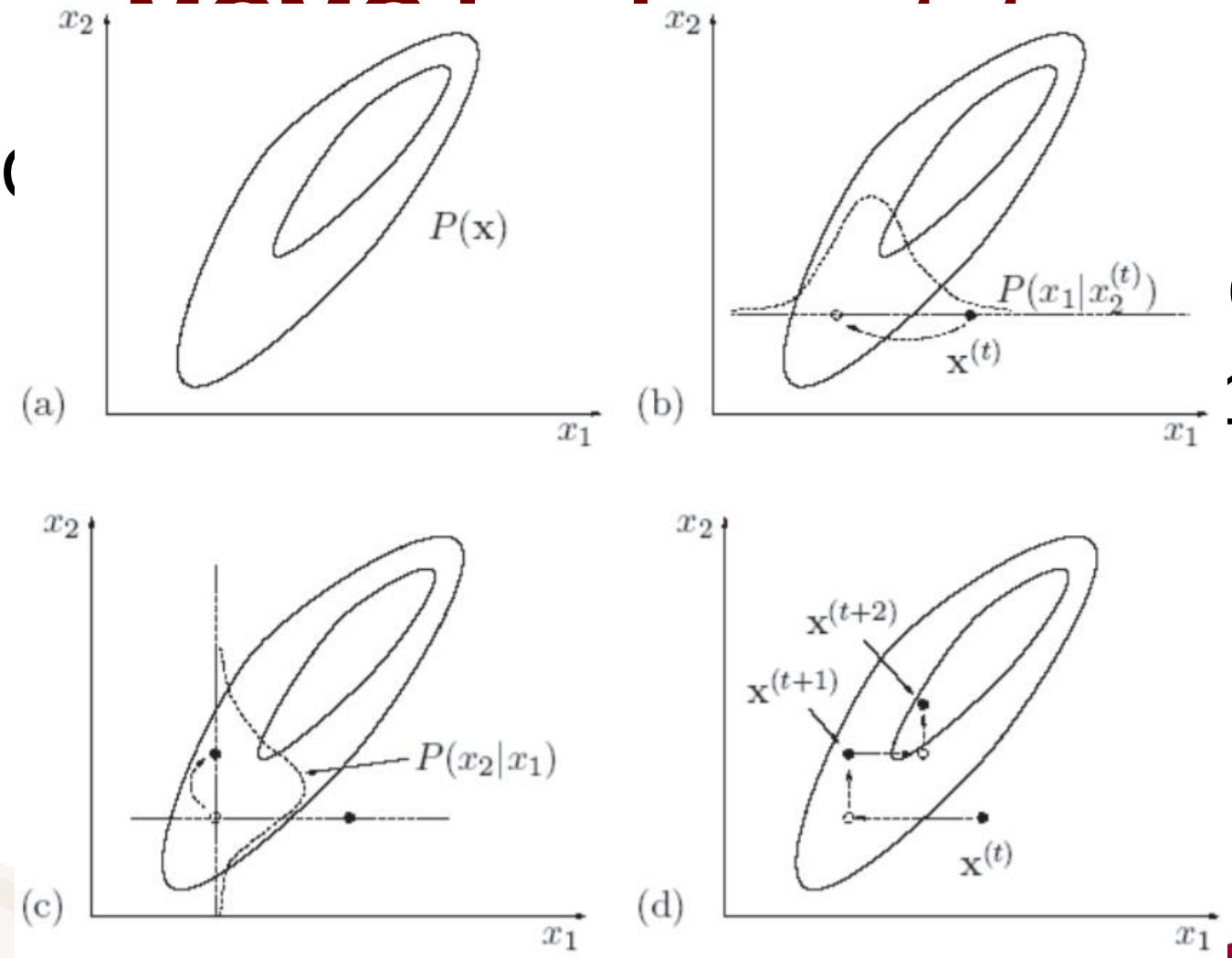
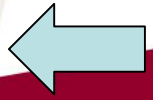


Figure 7: Example of Gibbs sampling in a 2D Gaussian. Source: [Mac03].

Reversible Jump JMCMC

- MH-within-Gibbs
  - Combining advantages of both implementations for different nature of parameters involved
- Infinite and finite mixtures
  - Infinite: component number is between 1 to infinity.
  - Finite: component number is either
    - Fixed
    - Variable: RJMCMC



Input: Data observations $\mathcal{X}$ and components number M
Output: AGM mixture parameter set Θ

1) Initialization
2) Step t: For $t = 1, \ldots$

Gibbs sampling part

a) Generate $Z^{(t)}$ from Eq. (3)
b) Compute $n_j^{(t)}$ from Eq. (7)
c) Generate $p_j^{(t)}$ from Eq. (6)

Metropolis-Hastings part

d) Sample $\xi_j^{(t)}$ $(\mu_j^{(t)}, \sigma_{lj}^{(t)}, \sigma_{rj}^{(t)})$ from Eqs. (8)
e) Compute acceptance ratio r from Eq. (9)
f) Generate $\alpha = min[1, r]$ and $u \sim U_{[0,1]}$
g) If $u \geq \alpha$ then $\xi^{(t)} = \xi^{(t-1)}$

RJMCMC part

h) Generate $u' \sim U_{[0,1]}$. If $b_m >= u'$, perform split or birth step, then calculate acceptance probability $\mathcal{A}$. If the step is accepted, set $m = m + 1$.
i) Generate $u' \sim U_{[0,1]}$. If $d_m >= u'$, perform merge or death step, then calculate acceptance probability $\mathcal{A}'$. If the step is accepted, set $m = m - 1$.

Intrusion Dectection

Dataset: NSL-KDD'99

(Available: http://www.unb.ca/cic/datasets/nsl.html)

CONFUSION MATRICES AND STATISTICS OF GMM AND AGM

GMM

	NF [a]	*F* [b]
NF	4238	7505
F	3397	10052

AGM

	NF	*F*
NF	2456	9278
F	582	12867

	GMM	*AGM*
Accuracy	53.39%	60.86%
Precision	36.09%	20.93%
False Positive Rate	42.75%	41.90%
False Negative Rate	44.49%	19.16%

[a]Non fault-prone, [b]Fault-prone.

Conclusion

- A fully Bayesian analysis based on reversible jump MCMC of AGM model
- Spam filtering using AGM model
- Future work: Feature reduction, data-oriented model adjustment

Thank You!