Information Theoretic Novelty Detection

Machine Learning Group

Department of Computer Science - The University of Sheffield

M. Filippone (m.filippone@dcs.shef.ac.uk) and G. Sanguinetti

July 7th, 2009

Outline of the talk

Novelty Detection

- The goal of novelty detection is to identify novelties/outliers in set of observations (datasets)
- Novelty/Outlier:
  "an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism."
  D. Hawkins

Novelty Detection

- The goal of novelty detection is to identify novelties/outliers in set of observations (datasets)
- Novelty/Outlier:
  "*an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism.*"
  D. Hawkins

Novelty Detection

Novelty/Outlier detection can be used for two different reasons:

- reduce their impact in the modeling stage (outlier detection)
- flag events/detect changes in order to take decisions on the system (novelty detection)

Novelty Detection

Novelty detection is employed in many fields:

- Mechanical Engineering (Fault detection)
- Condition Monitoring
- Hydrology

Novelty Detection

Two types of novelties:

- Event based (Additive Outliers)
- Model based (Innovative Outliers)

Novelty Detection

The performances of novelty detection systems can be measured by means of:

- Accuracy
- False Positive and False Negative rates

In every application it is important to understand what is the cost of False Negatives and False Positives.

Novelty Detection

- Neural networks
- Extreme value statistics
- Support Vector methods
- Statistical approaches (Frequentist and Bayesian)

Novelty Detection

- Modeling the system in a training stage
- Training set:

$$X = \{x_1, \ldots, x_n\}$$

- The model describes what is "normal", on the basis of the training set X
- No alternative hypothesis
- x_* will denote a test point

Maximum Likelihood Approach for i.i.d. data

- Assume a parametric form for $p(x)$, i.e. $p(x) = p(x|\theta)$
- The samples are i.i.d.
- Likelihood

$$L = \prod_{i=1}^{n} p(x_i|\theta)$$

- ML approach leads to an estimate $\hat{\theta}$ of θ
- x_* can be tested using quantiles

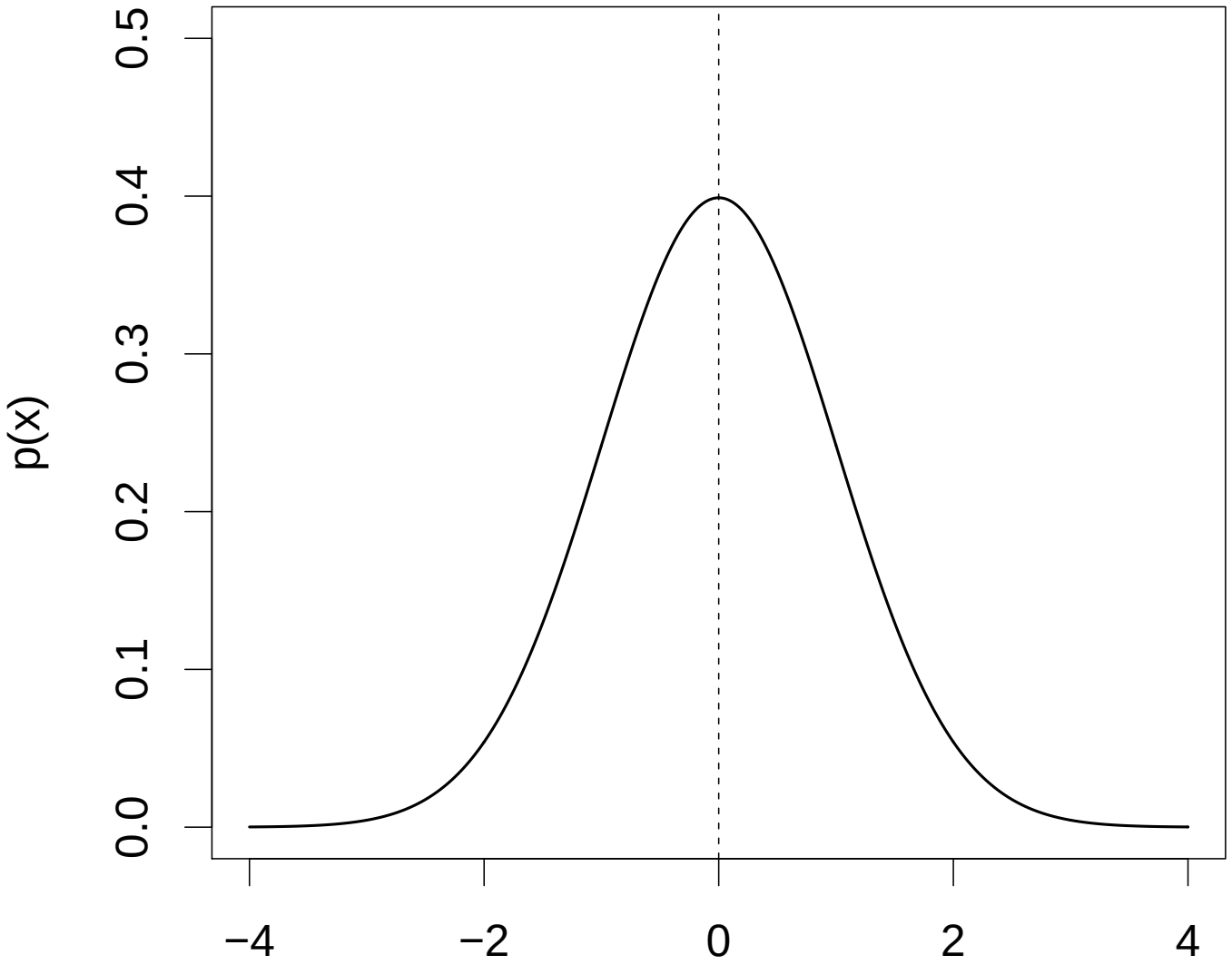
Maximum Likelihood Approach for i.i.d. data

- Assume a parametric form for $p(x)$, i.e. $p(x) = p(x|\theta)$
- The samples are i.i.d.
- Likelihood

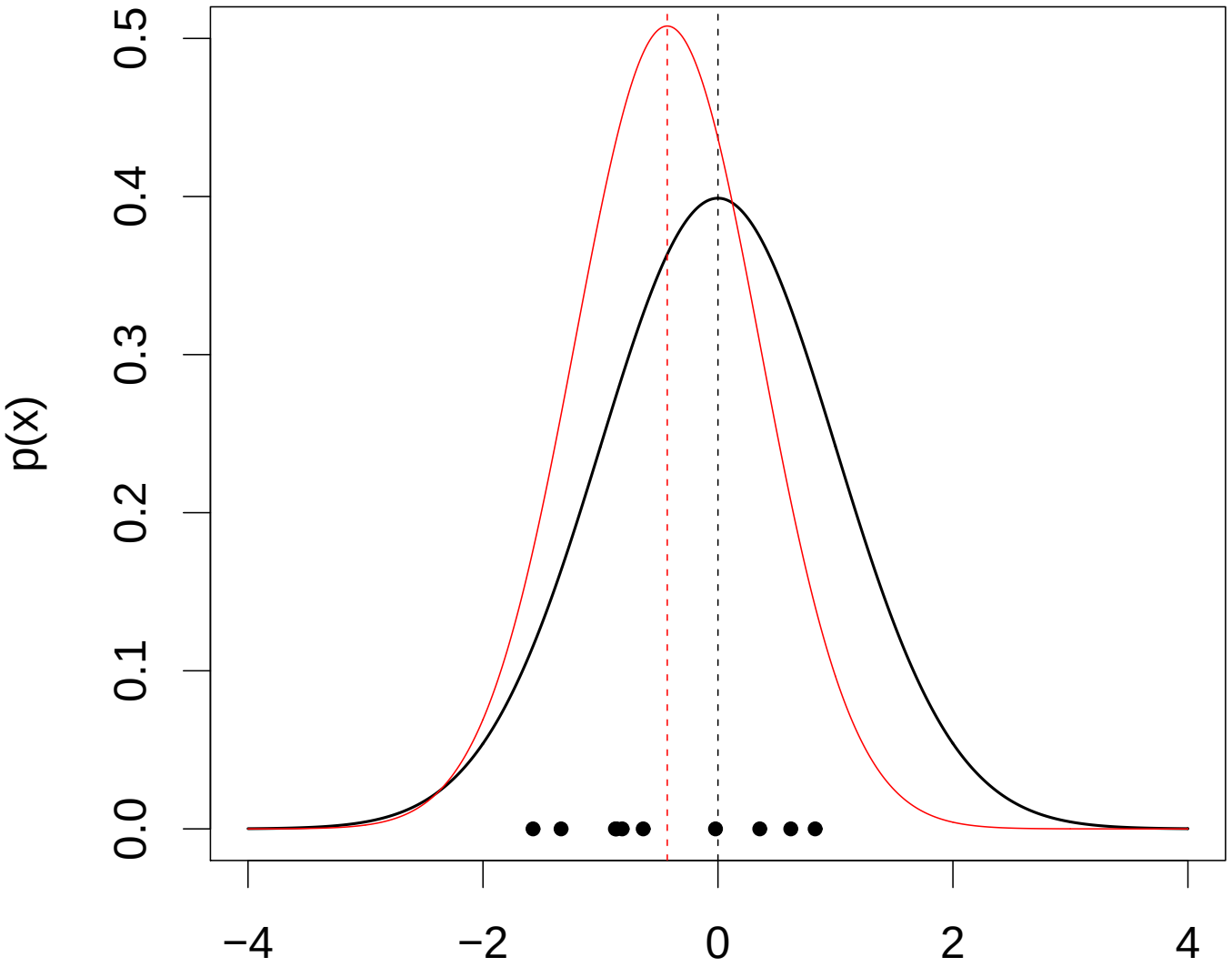
$$L = \prod_{i=1}^{n} p(x_i|\theta)$$

- ML approach leads to an estimate $\hat{\theta}$ of θ
- x_* can be tested using quantiles

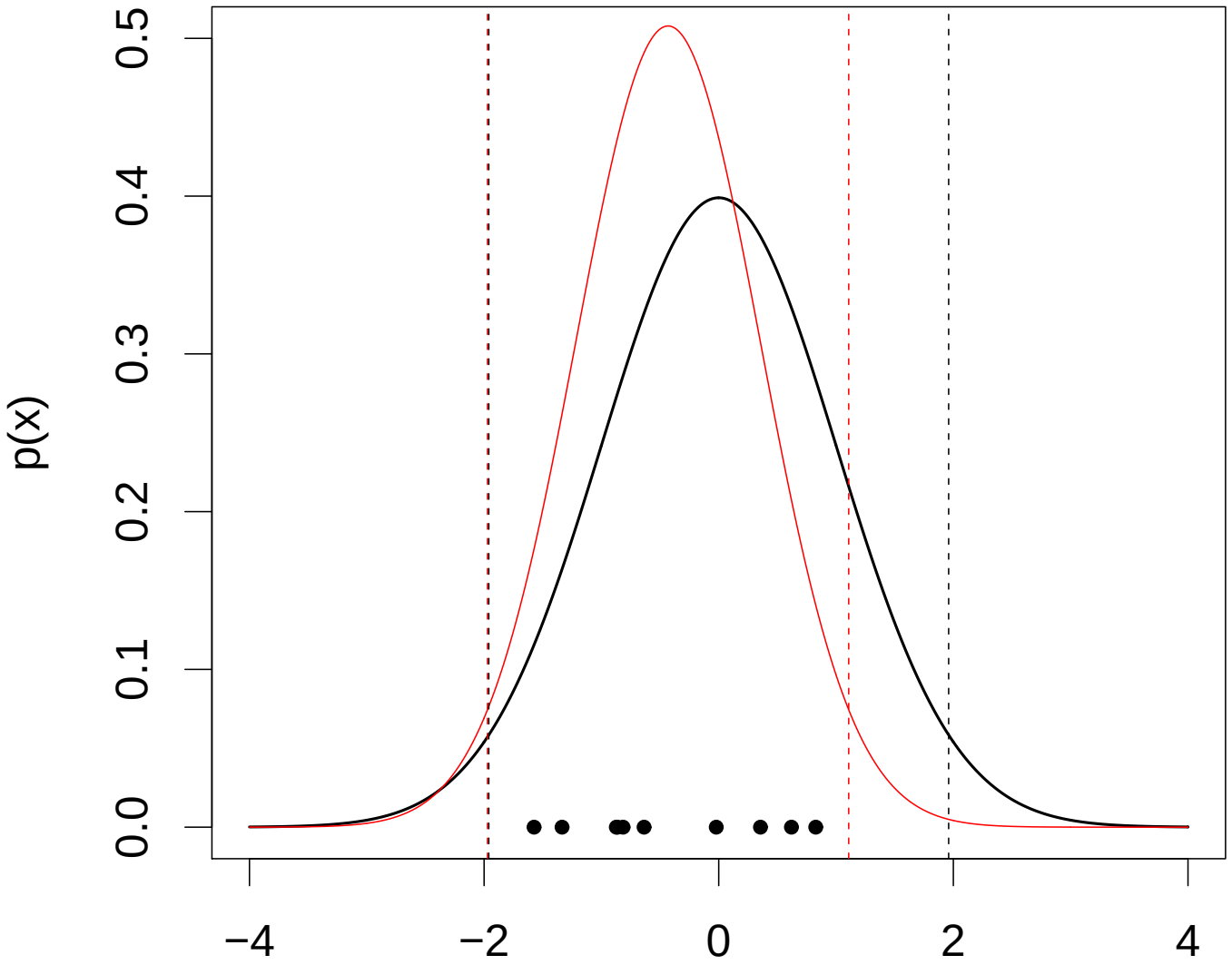
Example



Example



Example



Information Theoretic Novelty Detection

We recast the novelty detection problem in the framework of information theory

- i.i.d. data:
  - Gaussian case (univariate and multivariate)
  - Mixture of Gaussians (univariate and multivariate)
- time series (linear autoregressive)

Information Theoretic Novelty Detection for i.i.d. data

- $p(x|\hat{\theta})$ with $\hat{\theta}$ estimated on X
- $p(x|\hat{\theta}_*)$ with $\hat{\theta}_*$ estimated on $X \cup \{x_*\}$
- Kullback Leibler divergence between $p(x|\hat{\theta})$ and $p(x|\hat{\theta}_*)$
- KL divergence:

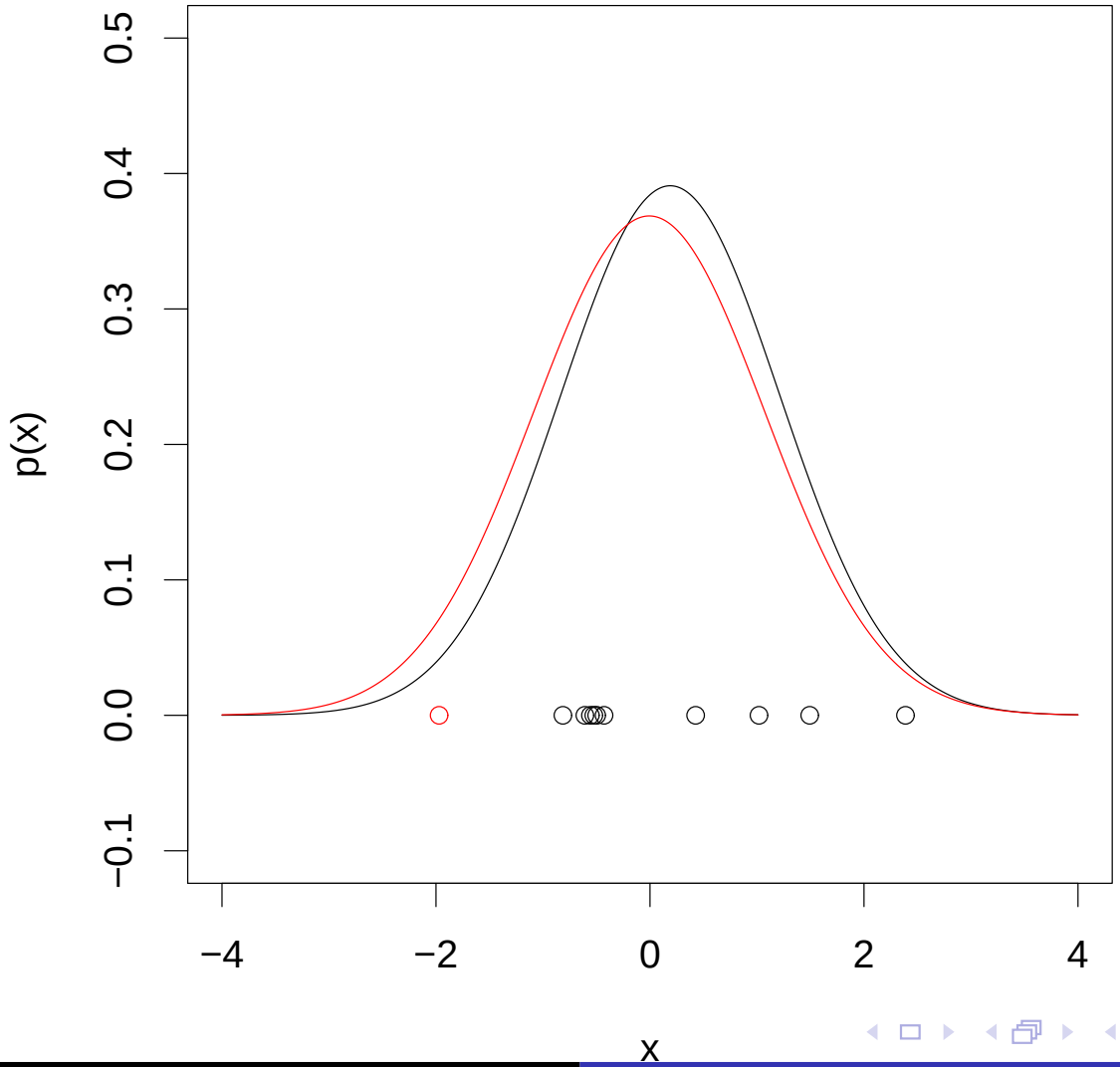
$$\mathrm{KL}\left[p\|q\right] = \int p(x) \log \left[\frac{p(x)}{q(x)}\right] dx$$

Information Theoretic Novelty Detection for i.i.d. data

- $p(x|\hat{\theta})$ with $\hat{\theta}$ estimated on X
- $p(x|\hat{\theta}_*)$ with $\hat{\theta}_*$ estimated on $X \cup \{x_*\}$
- Kullback Leibler divergence between $p(x|\hat{\theta})$ and $p(x|\hat{\theta}_*)$
- KL divergence:

$$\mathrm{KL}\left[p\|q\right] = \int p(x) \log \left[\frac{p(x)}{q(x)}\right] dx$$

KL divergence - Example



Univariate Gaussian Case

- $x_i \sim \mathcal{N}(m, s^2)$
- We introduce:

$$\hat{z} = \frac{(x_* - \hat{m})}{\hat{s}}$$

- The KL divergence results in:

$$\mathrm{KL} = f(n, \hat{z}^2)$$

- The distribution of $\hat{z}^2$ is known:

$$\hat{z}^2 = \frac{(x_* - \hat{m})^2}{\hat{s}^2} \sim \left(\frac{n+1}{n-1}\right) F_{(1,n-1)}$$

- The KL divergence is independent from the statistics!!

Univariate Gaussian Case

- $x_i \sim \mathcal{N}(m, s^2)$
- We introduce:

$$\hat{z} = \frac{(x_* - \hat{m})}{\hat{s}}$$

- The KL divergence results in:

$$\mathrm{KL} = f(n, \hat{z}^2)$$

- The distribution of $\hat{z}^2$ is known:

$$\hat{z}^2 = \frac{(x_* - \hat{m})^2}{\hat{s}^2} \sim \left(\frac{n+1}{n-1}\right) F_{(1,n-1)}$$

- The KL divergence is independent from the statistics!!

Univariate Gaussian Case

- Testing $\hat{z}^2$ or the KL divergence leads to the same results
- The thresholds for novelty can be set by using the quantiles of an $F_{(1,n-1)}$ with the desired different rejection rates
- x_* can be tested comparing $\hat{z}^2$ with the thresholds
- Since the distribution of $\hat{z}^2$ is exact, the test takes into account the variability due to the finite sample effect, and on average gives the expected false positive rate.

Univariate Gaussian Case

- Testing $\hat{z}^2$ or the KL divergence leads to the same results
- The thresholds for novelty can be set by using the quantiles of an $F_{(1,n-1)}$ with the desired different rejection rates
- x_* can be tested comparing $\hat{z}^2$ with the thresholds
- Since the distribution of $\hat{z}^2$ is exact, the test takes into account the variability due to the finite sample effect, and on average gives the expected false positive rate.

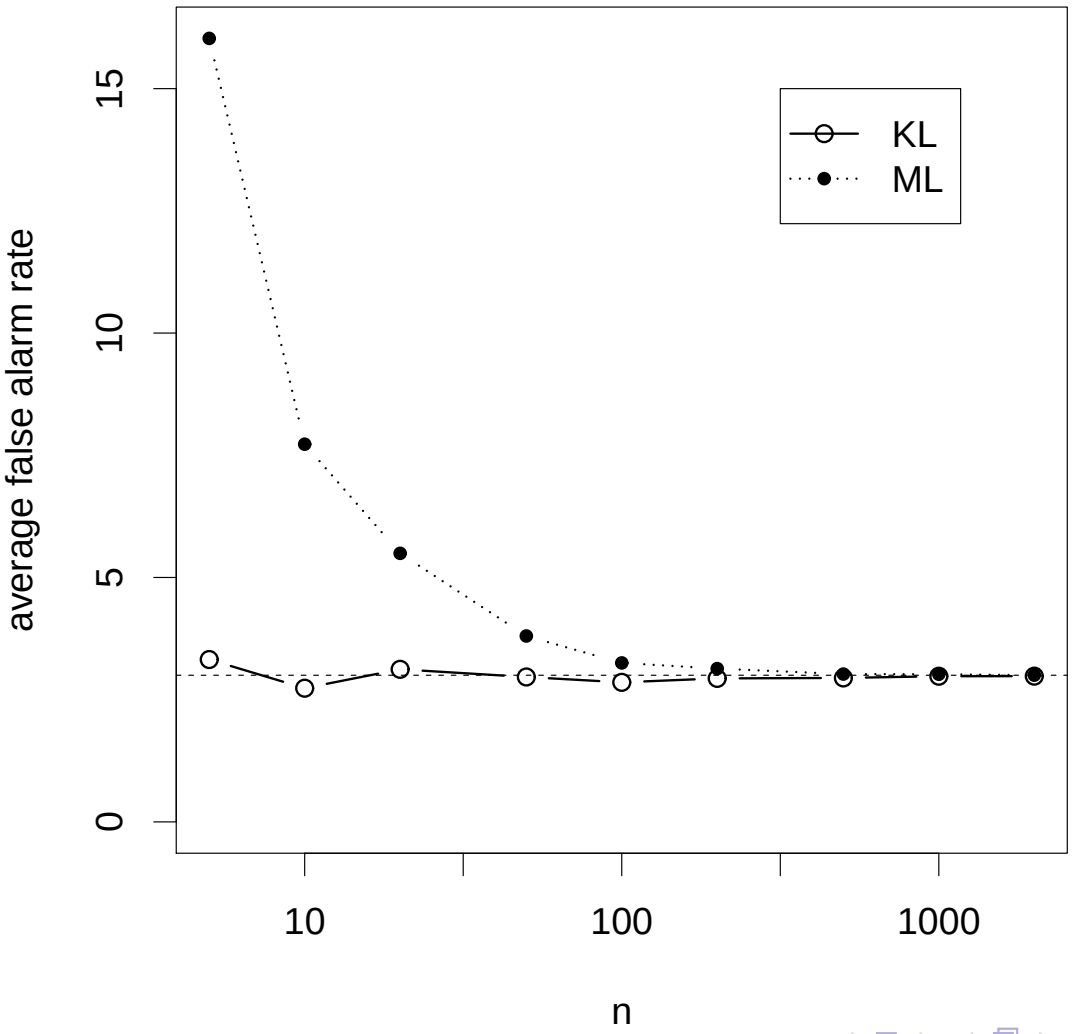
Univariate Gaussian Case

- Testing $\hat{z}^2$ or the KL divergence leads to the same results
- The thresholds for novelty can be set by using the quantiles of an $F_{(1,n-1)}$ with the desired different rejection rates
- x_* can be tested comparing $\hat{z}^2$ with the thresholds
- Since the distribution of $\hat{z}^2$ is exact, the test takes into account the variability due to the finite sample effect, and on average gives the expected false positive rate.

Univariate Gaussian Case - Experimental comparison

- Generate a training set of n points from a $\mathcal{N}(m, s^2)$;
- Generate 10^6 test points from the same $\mathcal{N}(m, s^2)$;
- Compute the number of outliers (false alarm rate);
- Repeat 200 times, and average the false alarm rate.

KL vs ML - Univariate Gaussian



Multivariate Gaussian Case

- Training data:

$$\mathbf{x}_i \sim \mathcal{N}(\mathbf{m}, S)$$

- Introduce:

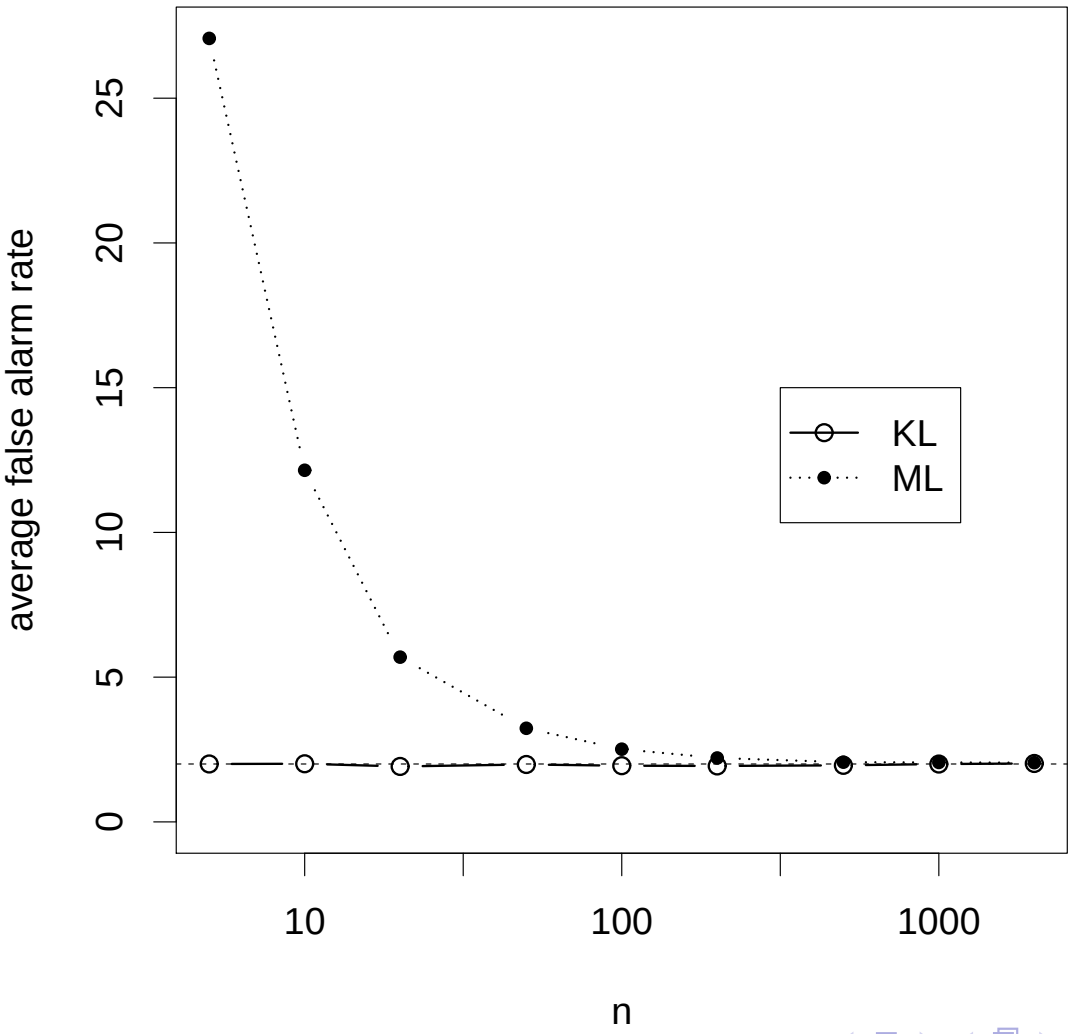
$$\hat{z}^2 = (\mathbf{x}_* - \hat{\mathbf{m}})^T \hat{S}^{-1} (\mathbf{x}_* - \hat{\mathbf{m}})$$

- The KL divergence results in:

$$\mathrm{KL} = f(n, \hat{z}^2)$$

- Again, the KL divergence does not depend on the statistics!!

KL vs ML - Multivariate Gaussian



Mixture of Gaussians

- Pdf:

$$p(x, \theta) = \sum_{k=1}^{c} \pi_k \mathcal{N}(x | m_k, s_k^2)$$

- KL divergence between:
  - $p(x|\hat{\theta})$ the mixture learned on X (for example using the EM algorithm)
  - $p(x|\hat{\theta}^*)$ the mixture learned starting from $p(x|\hat{\theta})$ and EM step on $X \cup \{x_*\}$
- No closed form for the KL divergence between two mixtures!!

Mixture of Gaussians

- Pdf:

$$p(x, \theta) = \sum_{k=1}^{c} \pi_k \mathcal{N}(x|m_k, s_k^2)$$

- KL divergence between:
  - $p(x|\hat{\theta})$ the mixture learned on X (for example using the EM algorithm)
  - $p(x|\hat{\theta}^*)$ the mixture learned starting from $p(x|\hat{\theta})$ and EM step on $X \cup \{x_*\}$
- No closed form for the KL divergence between two mixtures!!

Mixture of Gaussians

- Pdf:

$$p(x, \theta) = \sum_{k=1}^{c} \pi_k \mathcal{N}(x|m_k, s_k^2)$$

- KL divergence between:
  - $p(x|\hat{\theta})$ the mixture learned on X (for example using the EM algorithm)
  - $p(x|\hat{\theta}^*)$ the mixture learned starting from $p(x|\hat{\theta})$ and EM step on $X \cup \{x_*\}$
- No closed form for the KL divergence between two mixtures!!

Approximation of the KL divergence

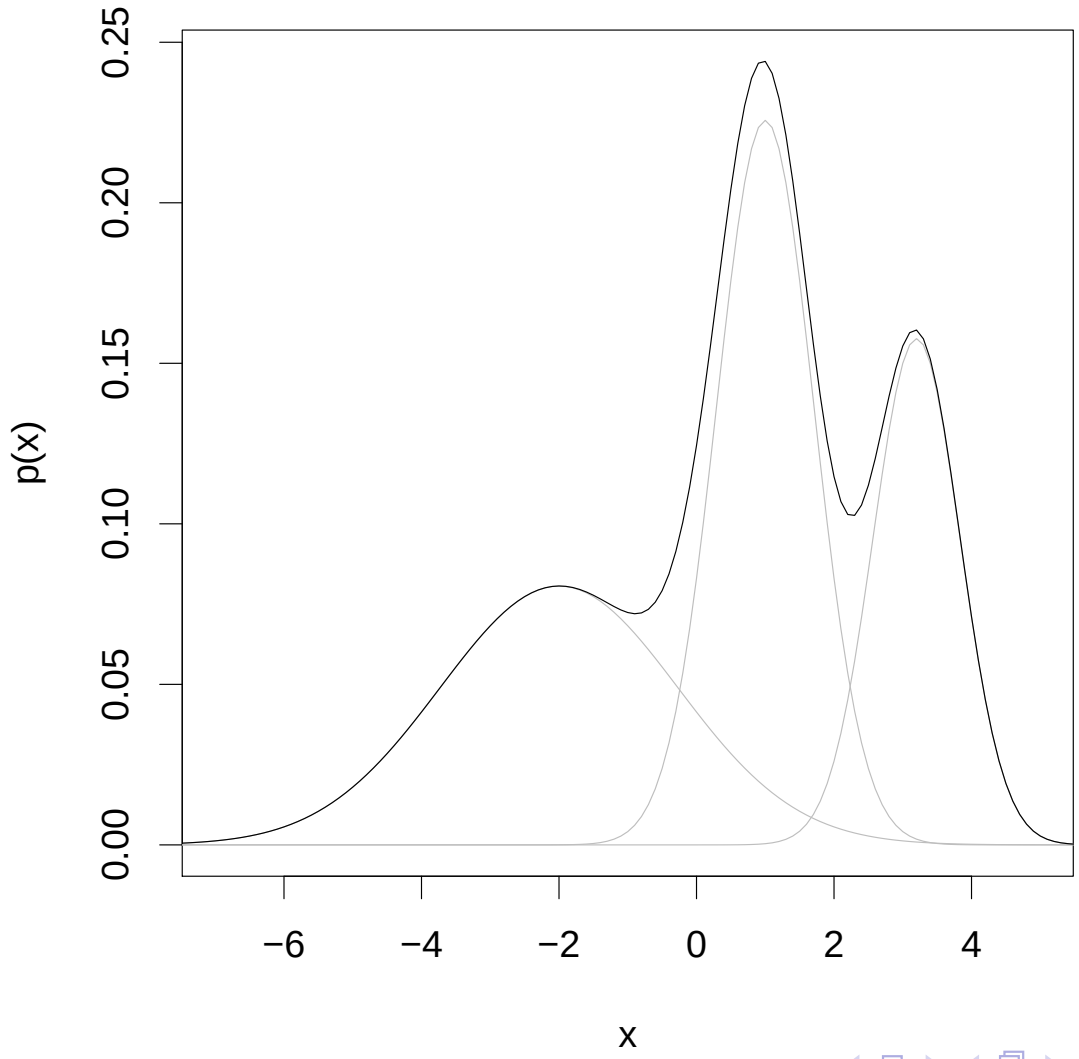
- Two-stage approximation
  - second order approximation of the logarithm

$$p(x|\hat{\theta}^*) = p(x|\hat{\theta}) + \delta p(x|\hat{\theta})$$

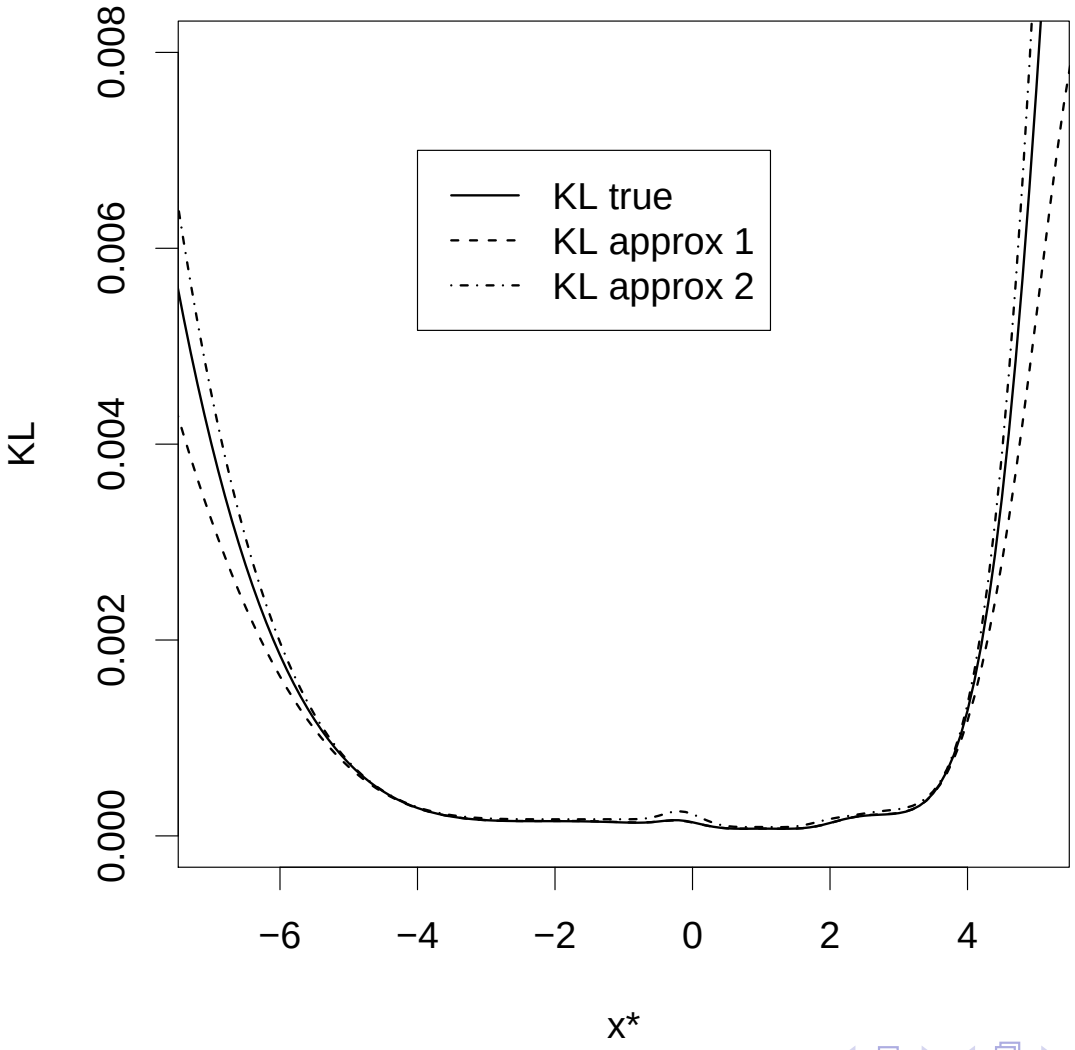
$$\log\left[\frac{p(x|\hat{\theta})}{p(x|\hat{\theta}^*)}\right] = -\log\left[1 + \frac{\delta p(x|\hat{\theta})}{p(x|\hat{\theta})}\right]$$

  - crisp responsibilities

Example - the pdf



Example - the approximation



Mixture of Gaussian - KL divergence

- the approximation of the KL divergence is:

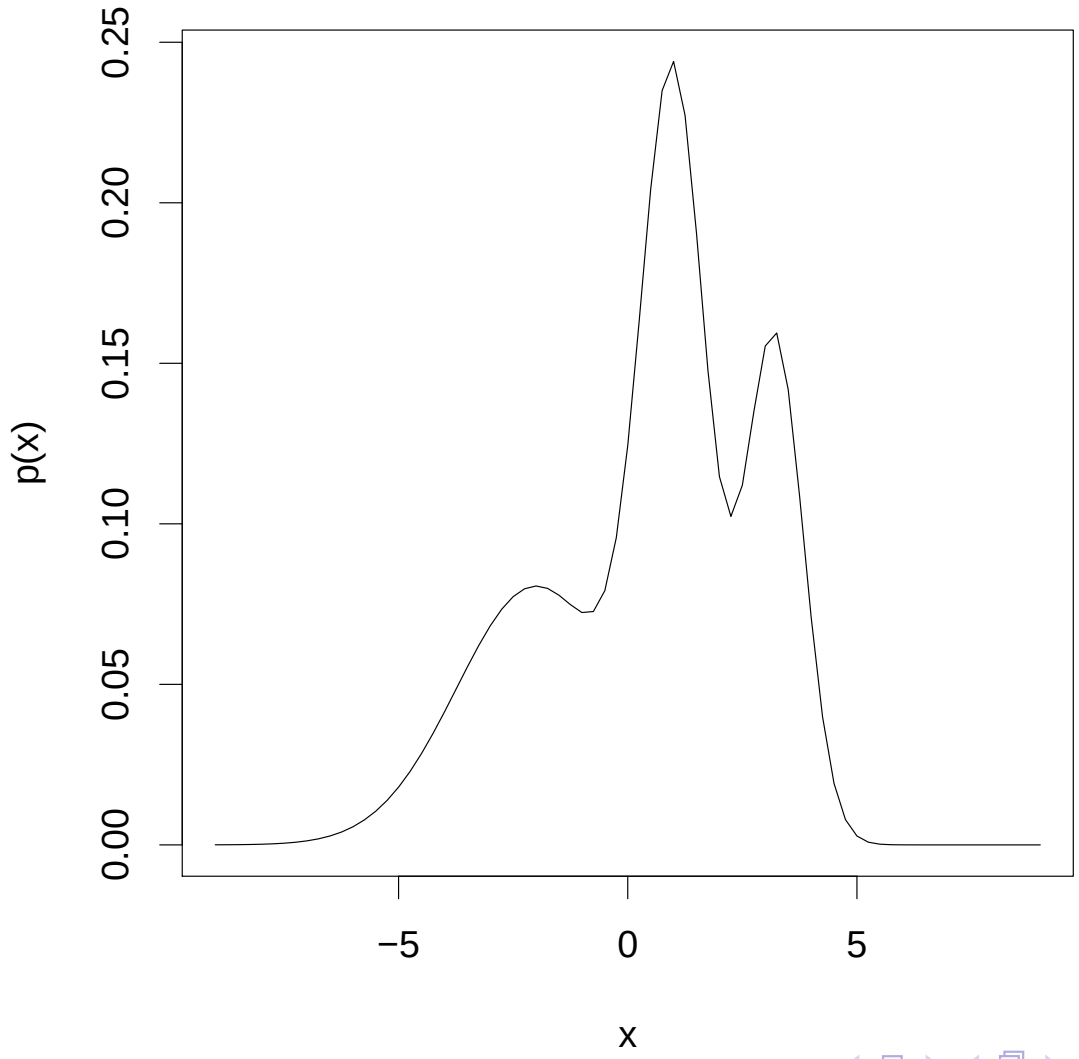
$$\mathrm{KL} = f(n, \hat{z}_k^2, \hat{\pi}_k, \hat{s}_k)$$

where:

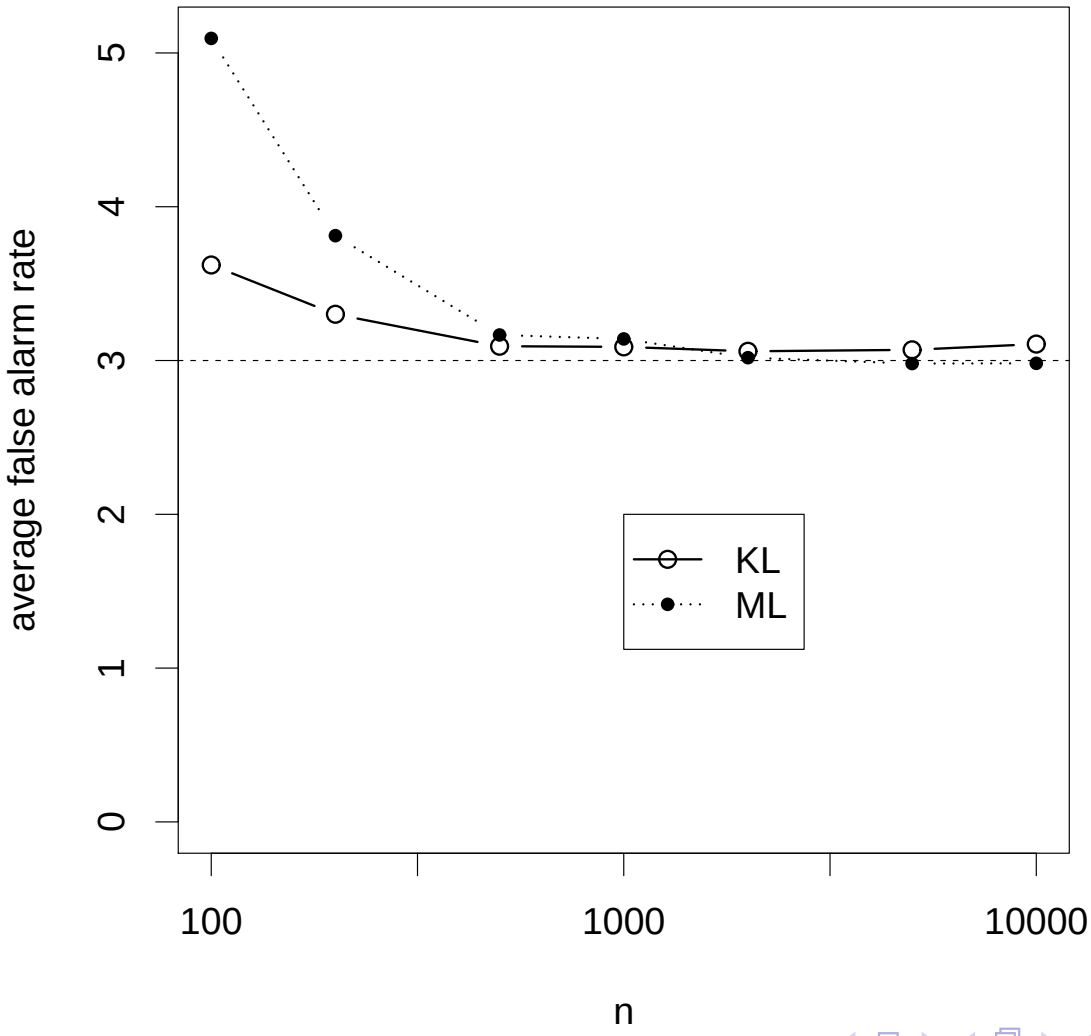
$$\hat{z}_k^2 = \frac{(x_* - \hat{m}_k)^2}{\hat{s}_k^2}$$

- Monte Carlo simulation to obtain the quantiles of the KL divergence
- We can take into account the variability of the means and the variances!!

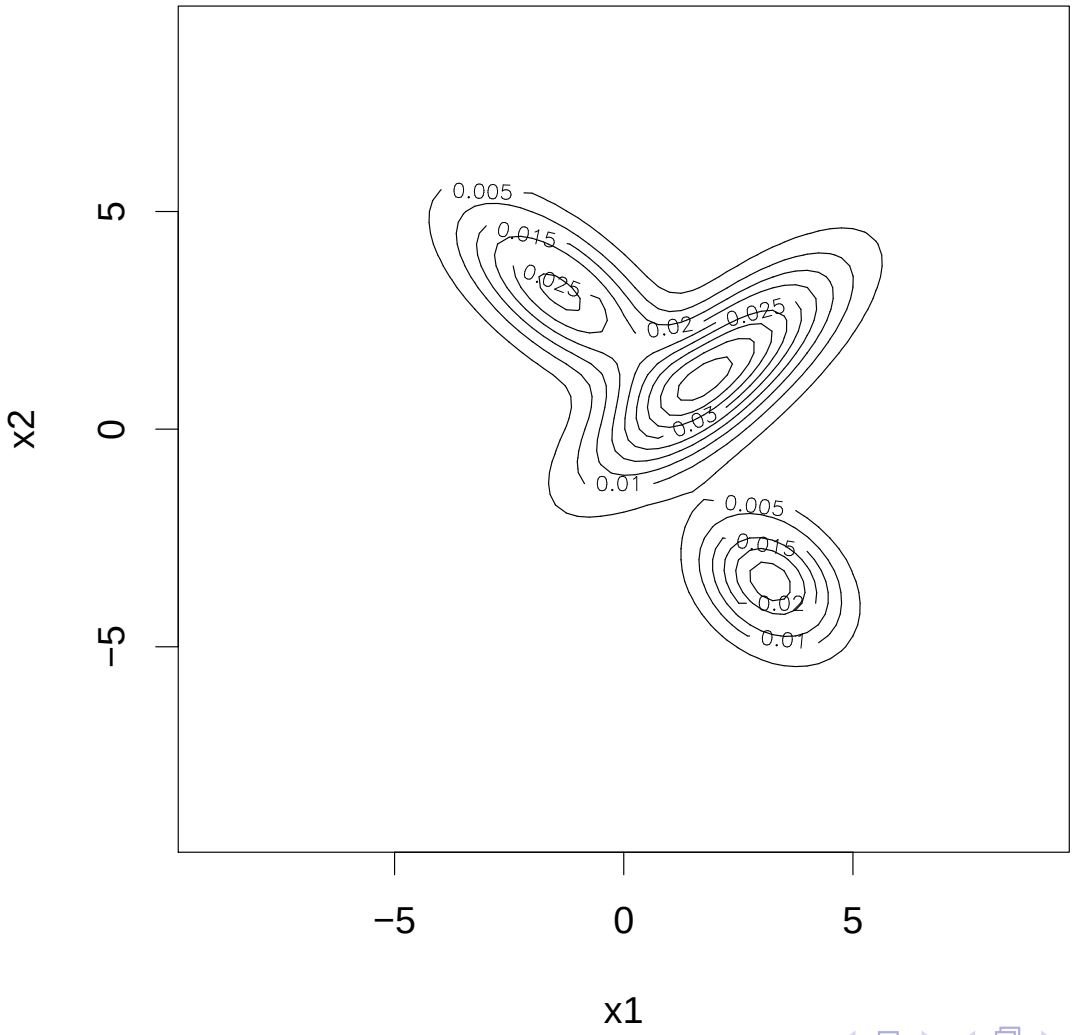
Mixture of Gaussian - Results



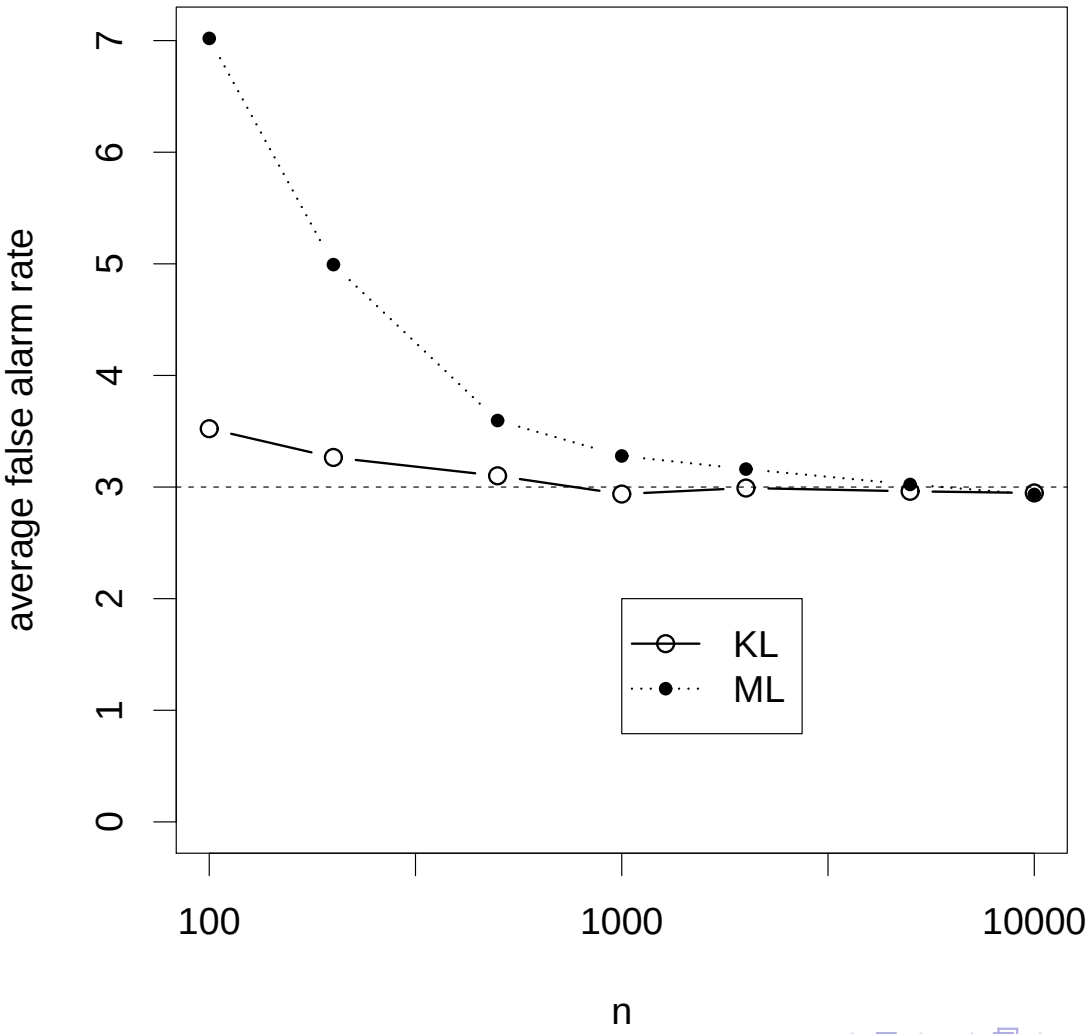
Mixture of Gaussian - Results



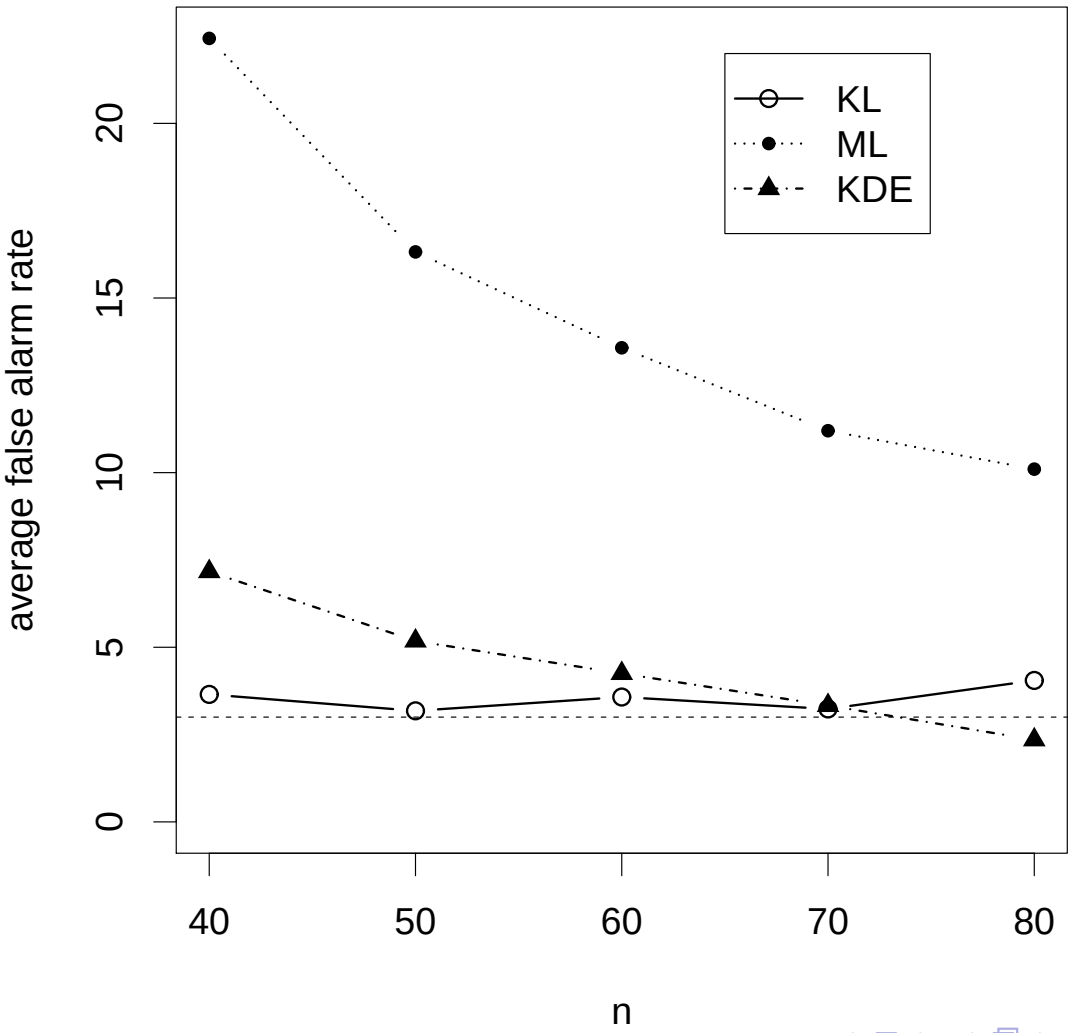
Mixture of Gaussian - Results



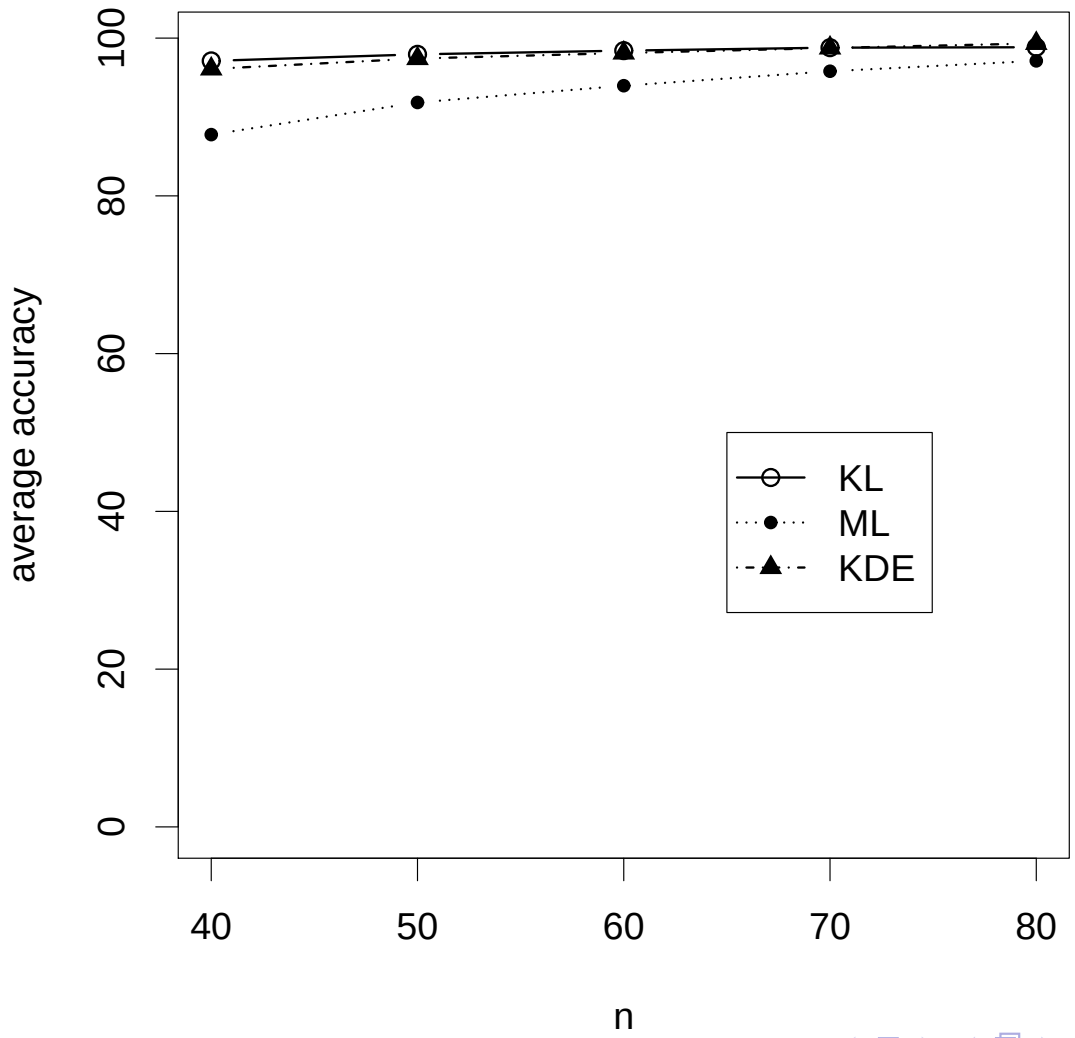
Mixture of Gaussian - Results



Iris - Results



Iris - Results



Autoregressive model - $\mathrm{AR}(d)$

- In many applications the i.i.d. assumption is not valid
- A well established framework for modeling temporal correlation in a series of observation is given by autoregressive models:

$$x_{t+1} = \boldsymbol{\alpha}^{\mathrm{T}} \mathbf{x}_t + \varepsilon_{t+1} + \mu$$

- $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d)$
- $\mathbf{x}_t = (x_t, x_{t-1}, \ldots, x_{t-d+1})$
- $\varepsilon_{t+1} \sim \mathcal{N}(0, \gamma^2)$ and i.i.d.
- μ allows to model series with any mean value m

Autoregressive model - Parameter Estimation

The parameters of an $\mathrm{AR}(d)$ can be estimated in many ways

- Yule-Walker method
- Variational Bayes
- Spectral techniques
- Maximum Likelihood
- Kalman filter

Autoregressive model - Parameter Estimation

$$c_k = \mathrm{E}[(x_i - m)(x_{i-k} - m)] \qquad k = 1, \ldots, d$$

Introducing the vector $\mathbf{c} = (c_1, c_2, \ldots, c_d)^{\mathrm{T}}$ and the matrix C:

$$C = \begin{pmatrix} c_0 & c_1 & \cdots & c_{d-1} \\ c_1 & c_0 & \cdots & c_{d-2} \\ \vdots & \vdots & \ddots & \vdots \\ c_{d-1} & c_{d-2} & \cdots & c_0 \end{pmatrix}$$

we see that:

$$\boldsymbol{\alpha} = C^{-1}\mathbf{c}$$

Autoregressive model - Parameter Estimation

Once we have $\hat{\boldsymbol{\alpha}}$, we can estimate the other parameters of the model μ and γ.
Let's focus on $\hat{\gamma}^2$

$$\hat{\gamma}^2 = \frac{1}{n-d}\sum_{i=d}^{n-1}\left(x_{i+1} - \hat{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{x}_i - \hat{\mu}\right)^2 = \frac{1}{n-d}\sum_{i=d}^{n-1}\hat{\varepsilon}_{i+1}^2$$

In a ML approach to novelty detection we test a new data point on the basis of $\hat{\gamma}^2$

Autoregressive model - Information theoretic measure

- Updated version of the parameters when we add a new data point x_*: $\hat{\boldsymbol{\alpha}}_*$, $\hat{\mu}_*$, and $\hat{\gamma}_*^2$
- Information content of x_* in the null hypothesis that it has been generated from the same model:

$$\mathrm{KL}\left[\mathcal{N}(\varepsilon|0,\hat{\gamma}^2)\|\mathcal{N}(\varepsilon|0,\hat{\gamma}_*^2)\right] = f\left(\frac{\hat{\gamma}_*^2}{\hat{\gamma}^2}\right)$$

Approximating the KL divergence

Let's focus on the ratio $\frac{\hat{\gamma}_*^2}{\hat{\gamma}^2}$

- Write the estimated parameters as their true values plus a term that is given by the fact that the estimation is based on a finite set of observations. For $\boldsymbol{\alpha}$, for example:

$$\hat{\boldsymbol{\alpha}} = \boldsymbol{\alpha} + \Delta\boldsymbol{\alpha} \qquad \hat{\boldsymbol{\alpha}}_* = \boldsymbol{\alpha} + \Delta\boldsymbol{\alpha}_*$$

- Substitute these relations into $\hat{\gamma}^2$ and $\hat{\gamma}_*^2$
- Compute a first order expansion of $\frac{\hat{\gamma}_*^2}{\hat{\gamma}^2}$

Approximating the KL divergence

Let's focus on the ratio $\frac{\hat{\gamma}_*^2}{\hat{\gamma}^2}$

- Write the estimated parameters as their true values plus a term that is given by the fact that the estimation is based on a finite set of observations. For $\boldsymbol{\alpha}$, for example:

$$\hat{\boldsymbol{\alpha}} = \boldsymbol{\alpha} + \Delta\boldsymbol{\alpha} \qquad \hat{\boldsymbol{\alpha}}_* = \boldsymbol{\alpha} + \Delta\boldsymbol{\alpha}_*$$

- Substitute these relations into $\hat{\gamma}^2$ and $\hat{\gamma}_*^2$
- Compute a first order expansion of $\frac{\hat{\gamma}_*^2}{\hat{\gamma}^2}$

Approximating the KL divergence

The ratio becomes a function of this form:

$$\frac{\hat{\gamma}_*^2}{\hat{\gamma}^2} \simeq \frac{n-d}{n-d+1}\left[1+\frac{\Delta}{\sum_{i=d}^{n-1}\varepsilon_{i+1}^2}\right]$$

where:

$$\Delta = \varepsilon_*^2 + \text{correction terms}$$

Approximating the KL divergence

- The leading term of the ratio $\frac{\Delta}{\sum_{i=d}^{n-1} \varepsilon_{i+1}^2}$ is therefore:

$$\frac{\varepsilon_*^2}{\sum_{i=d}^{n-1} \varepsilon_{i+1}^2} \sim \frac{1}{n-d} F_{(1,n-d)}$$

- We propose this approximation:

$$\frac{\Delta}{\sum_{i=d}^{n-1} \varepsilon_{i+1}^2} \sim \tau \frac{1}{n-d} F_{(1,n-d)}$$

- We compute τ to match the expected value of the F-distribution with the actual distribution of the ratio $\frac{\Delta}{\sum_{i=d}^{n-1} \varepsilon_{i+1}^2}$

Approximating the KL divergence

- The leading term of the ratio $\frac{\Delta}{\sum_{i=d}^{n-1} \varepsilon_{i+1}^2}$ is therefore:

$$\frac{\varepsilon_*^2}{\sum_{i=d}^{n-1} \varepsilon_{i+1}^2} \sim \frac{1}{n-d} F_{(1,n-d)}$$

- We propose this approximation:

$$\frac{\Delta}{\sum_{i=d}^{n-1} \varepsilon_{i+1}^2} \sim \tau \frac{1}{n-d} F_{(1,n-d)}$$

- We compute τ to match the expected value of the F-distribution with the actual distribution of the ratio $\frac{\Delta}{\sum_{i=d}^{n-1} \varepsilon_{i+1}^2}$

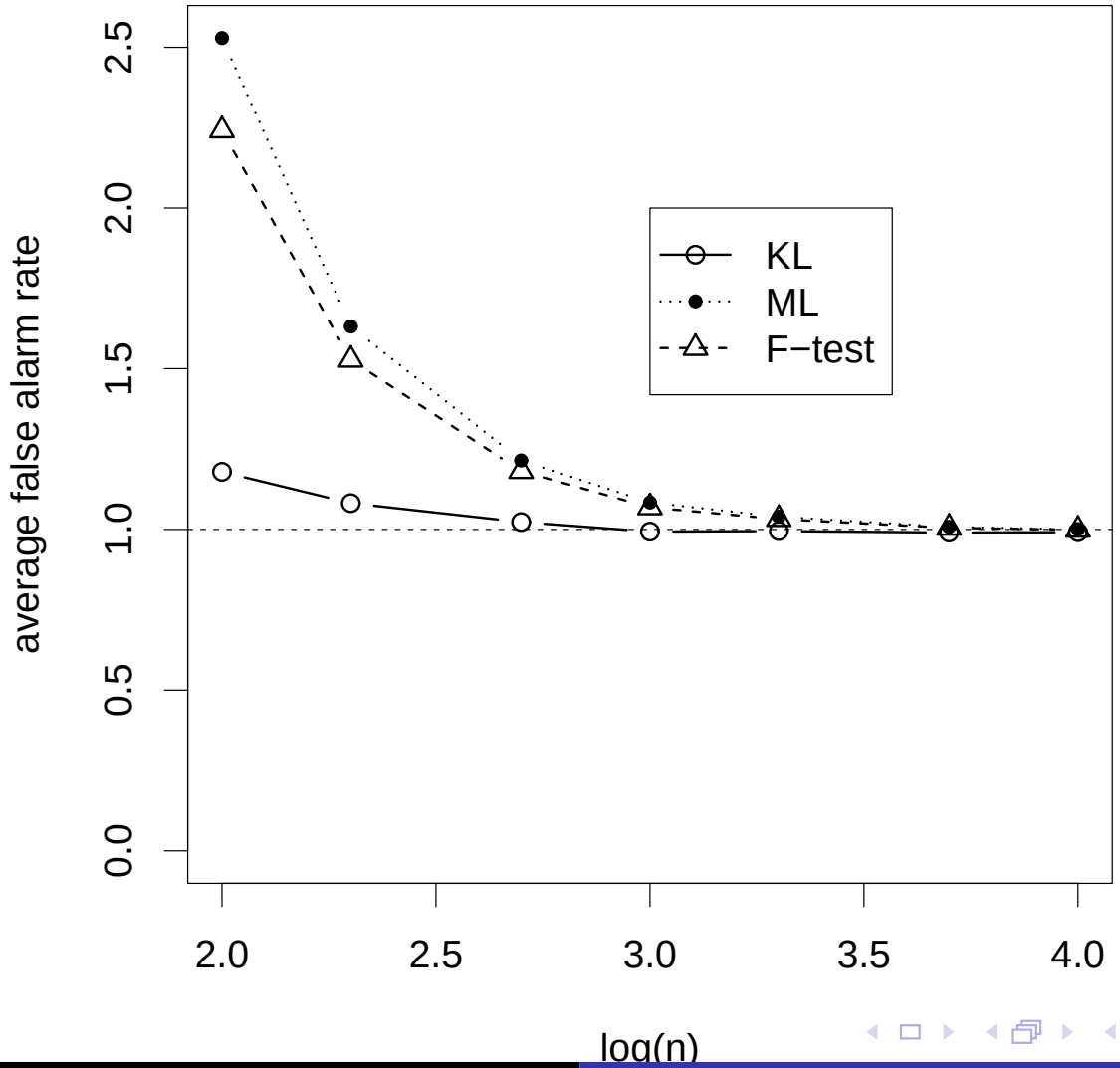
Approximating the KL divergence

Finally, the test we propose is:

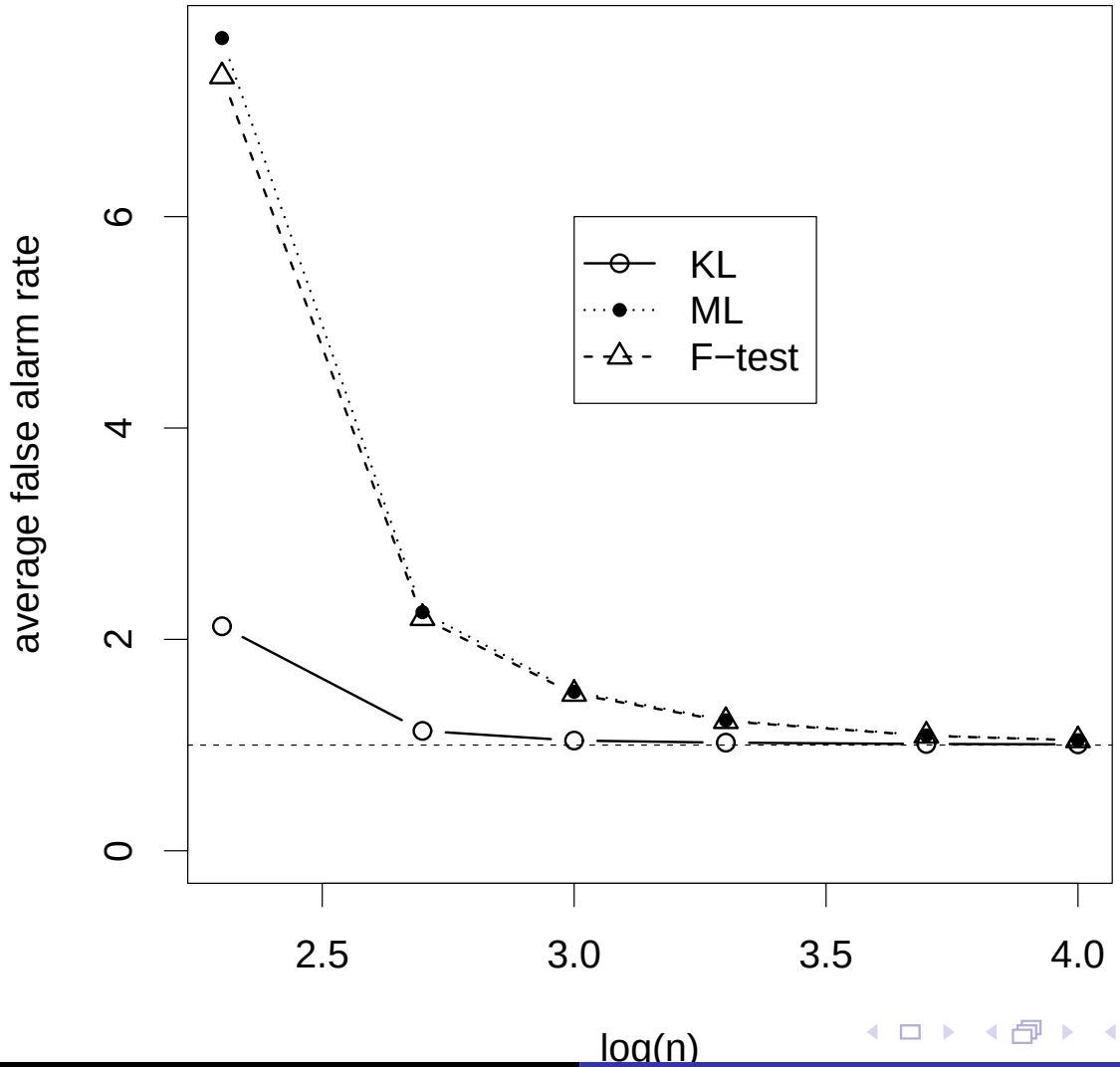
$$\frac{\hat{\gamma}_*^2}{\hat{\gamma}^2} = \frac{n-d}{n-d+1}\left[1+\tau\frac{1}{n-d}F_{(1,n-d)}\right]$$

$$\tau = \left(1+2\frac{d}{(n-d)}+\frac{2}{n}(1-\sum_i \hat{\alpha}_i)^2-\frac{2}{n}(1-\sum_i \hat{\alpha}_i)\right)$$

AR(10) - Results



AR(50) - Results



Conclusions and Future Works

- We recast novelty detection in the framework of information theory
- Important connections with statistical testing
- Control of the false positive rate even for small data sets
- Model selection is crucial
- Extension to the exponential family (?)
- Regularization (?)
- Extend to model based novelties

Conclusions and Future Works

- We recast novelty detection in the framework of information theory
- Important connections with statistical testing
- Control of the false positive rate even for small data sets
- Model selection is crucial
- Extension to the exponential family (?)
- Regularization (?)
- Extend to model based novelties

Conclusions and Future Works

- We recast novelty detection in the framework of information theory
- Important connections with statistical testing
- Control of the false positive rate even for small data sets
- Model selection is crucial
- Extension to the exponential family (?)
- Regularization (?)
- Extend to model based novelties

Thank You!