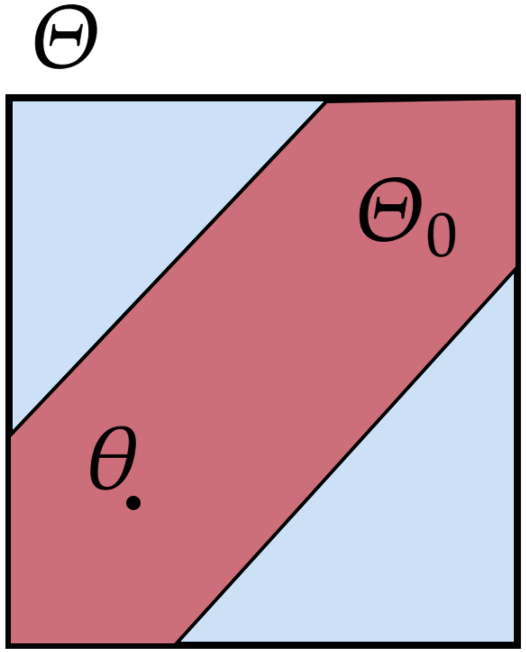
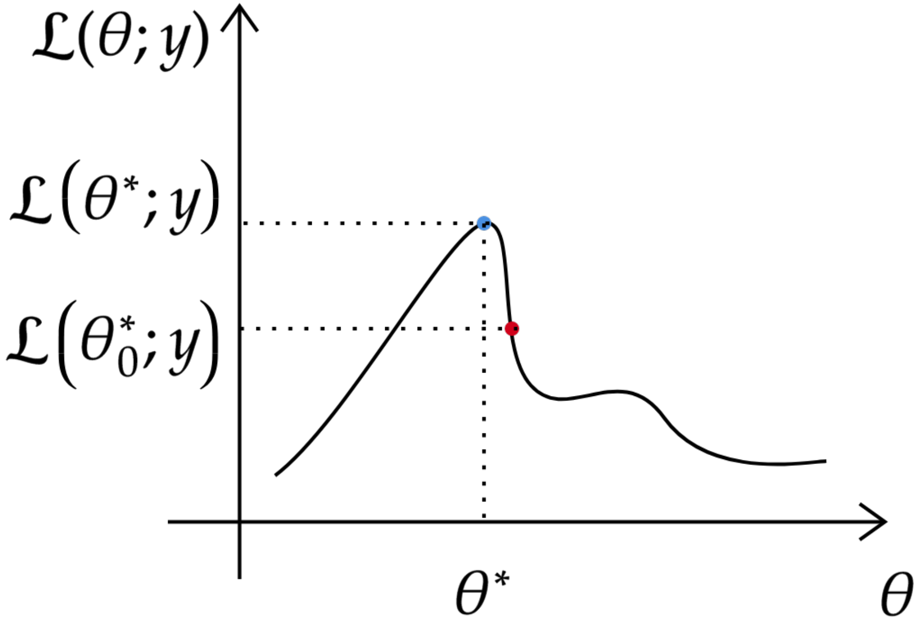
Likelihood Ratios, Derived Tests, and Applications

Alex Nguyen-Le

Intuition and Interpretation



$\Theta_0 \subset \Theta$



$$H_0 : \theta \in \Theta_0$$

$$H_1 : \theta \in \Theta \backslash \Theta_0$$

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \mathcal{L}(\theta; y)$$

$$\hat{\theta}_0 = \arg\max_{\theta \in \Theta} \mathcal{L}(\theta; y)$$

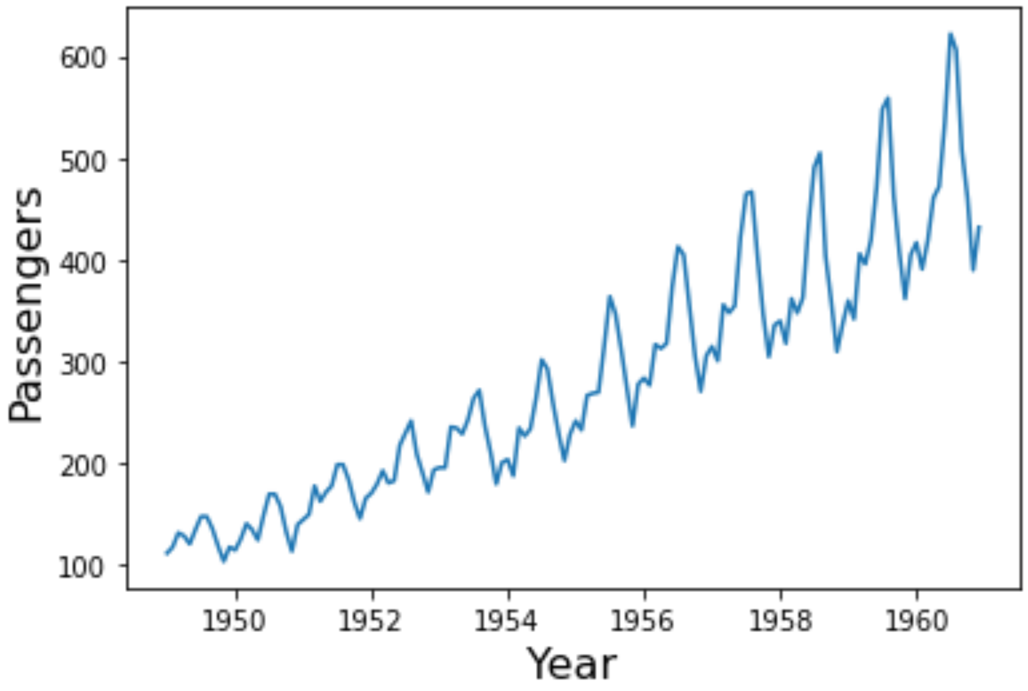
Likelihood Ratios for Model Selection

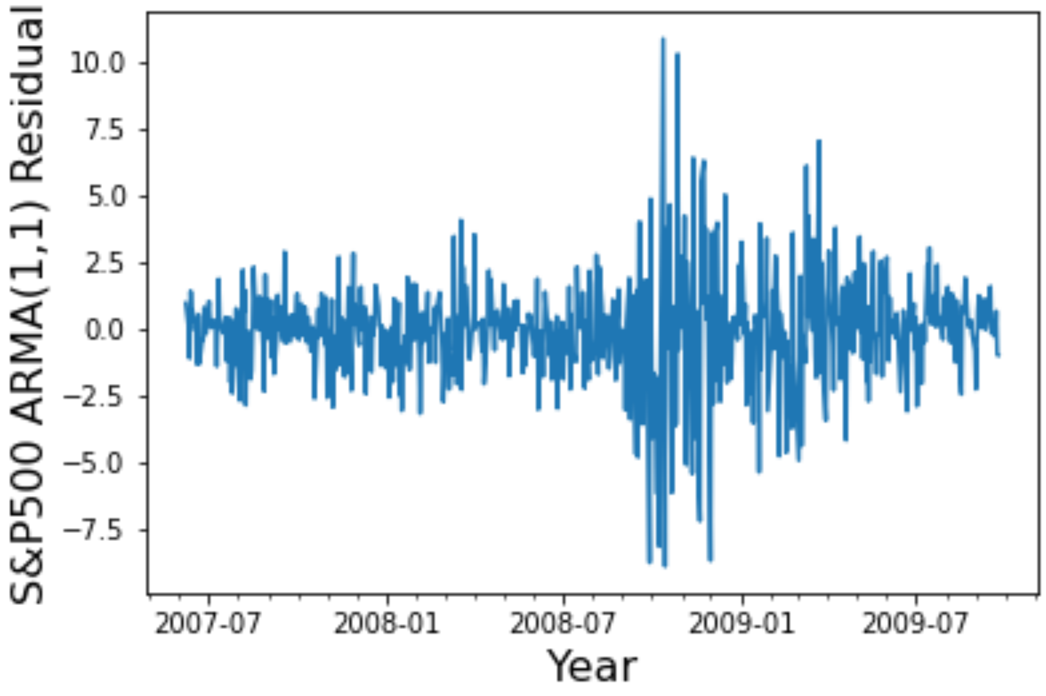
$$\mathrm{LR} \quad := \quad -2\ \log\left(\frac{\sup\limits_{\theta\in\Theta_0} p(\theta; y)}{\sup\limits_{\theta\in\Theta} p(\theta; y)}\right) \quad = \quad -2\ \left(\mathcal{L}(\hat{\theta}_0; y) - \mathcal{L}(\hat{\theta}; y)\right)$$

Typically, Θ_0 is a "submodel" constraint, e.g., some parameters are subject to equality contraints that simplify the model. This condition is also known as the nested model constraint.

Prototypical Applications

- Time series analysis
  - Is there a nonstationary mean?
  - Is there a GARCH component?





Wilk's Theorem and Asymptotic Results

$$\text{LR} \quad := \quad -2\left(\mathcal{L}(\hat{\theta}_0; y) - \mathcal{L}(\hat{\theta}; y)\right)$$

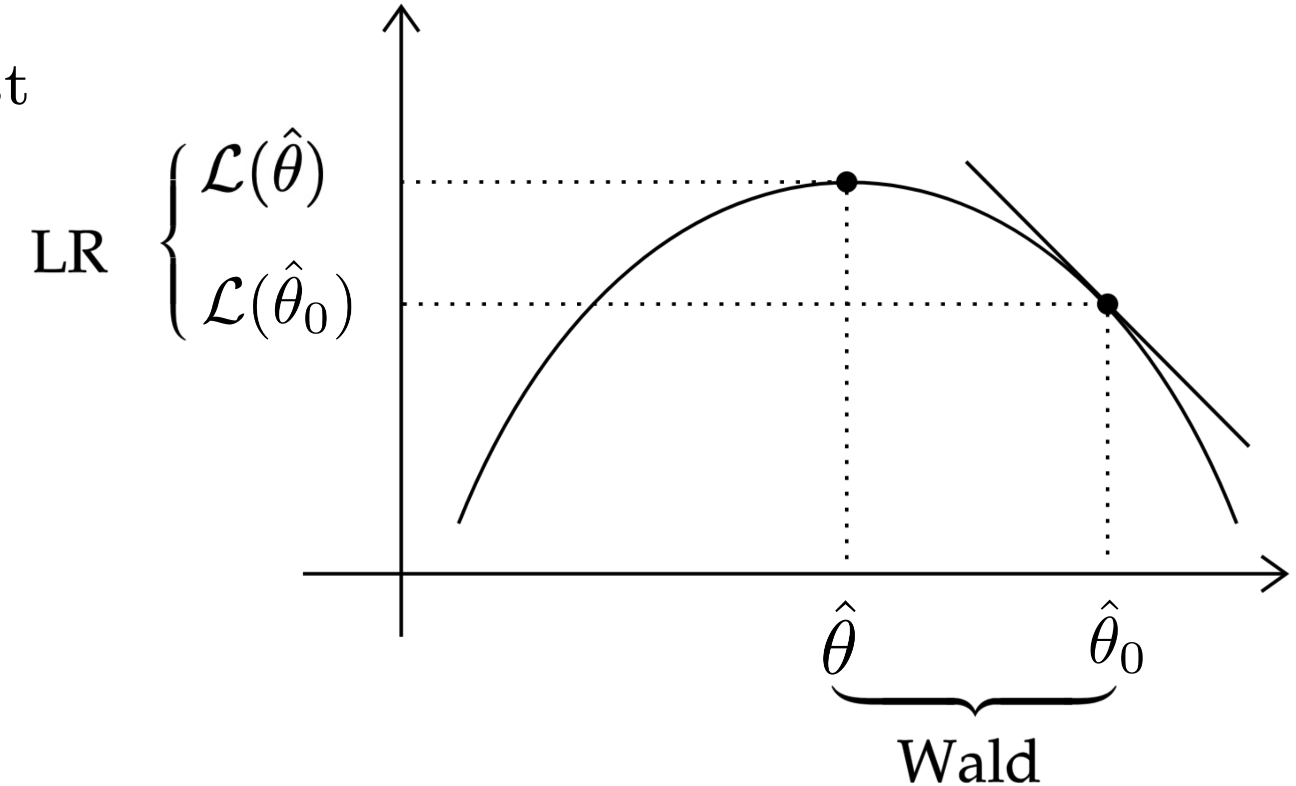
Wilk's Theorem (Informal)

Let θ^* satisfy the first order conditions for optimality, let θ^* converge in distribution to a normal, and let the ML Fisher Information matrix, $\mathcal{I}(\theta)$ be consistently estimated by $\mathcal{I}(\theta^*)$. Under the null hypothesis, the likelihood ratio statistic converges in distribution to χ^2 distribution with degrees of freedom equal to the number of equality constraints.

Approximations to the Likelihood Ratio

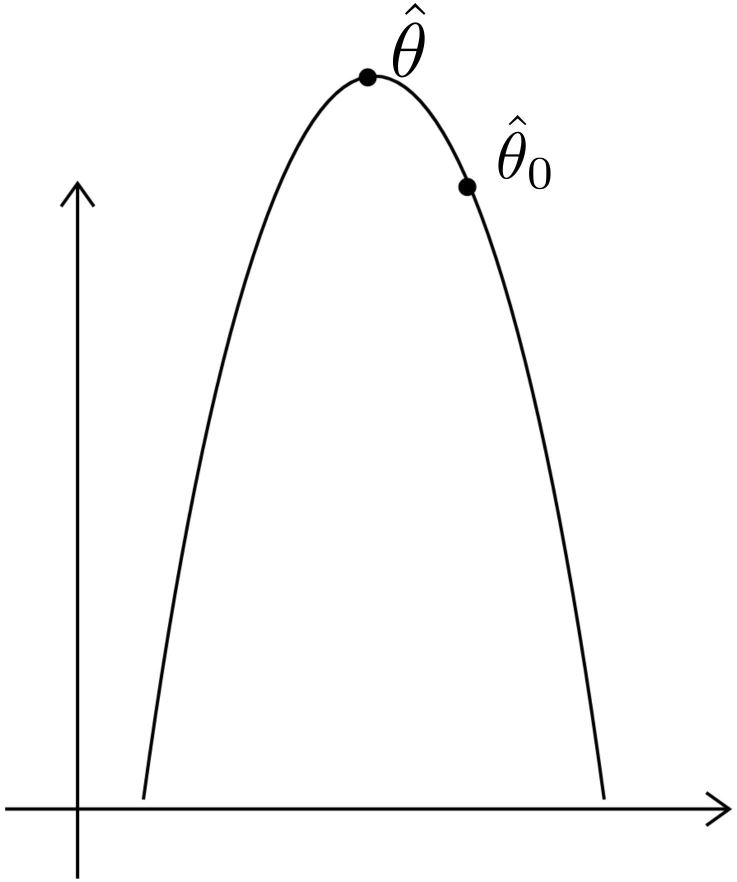
- Oftentimes, one of the optimization problems is much easier to solve!
  - Nested submodel constraint typically eliminates some model components

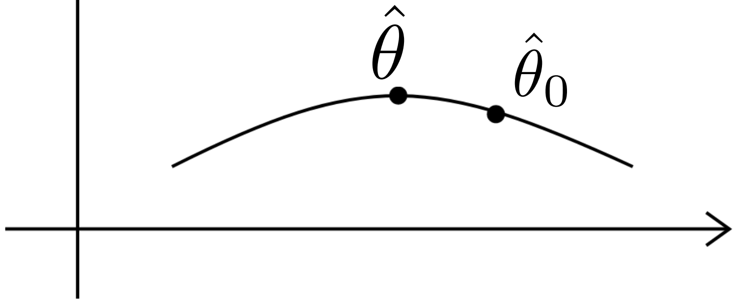
$$-2\,\log\left(\overbrace{\underbrace{\frac{\sup\limits_{\theta\in\Theta_0} p(\theta;y)}{\sup\limits_{\theta\in\Theta} p(\theta;y)}}_{\text{Wald Test}}}^{\text{Lagrange Multiplier Test}}\right)$$



Wald Test

- Key idea: distance between coordinates needs a correction that depends upon local curvature





Wald Test and Asymptotic results

$$W = (\hat{\theta} - \hat{\theta}_0)^{\mathsf{T}} \mathcal{I}(\hat{\theta}_0)(\hat{\theta} - \hat{\theta}_0)$$

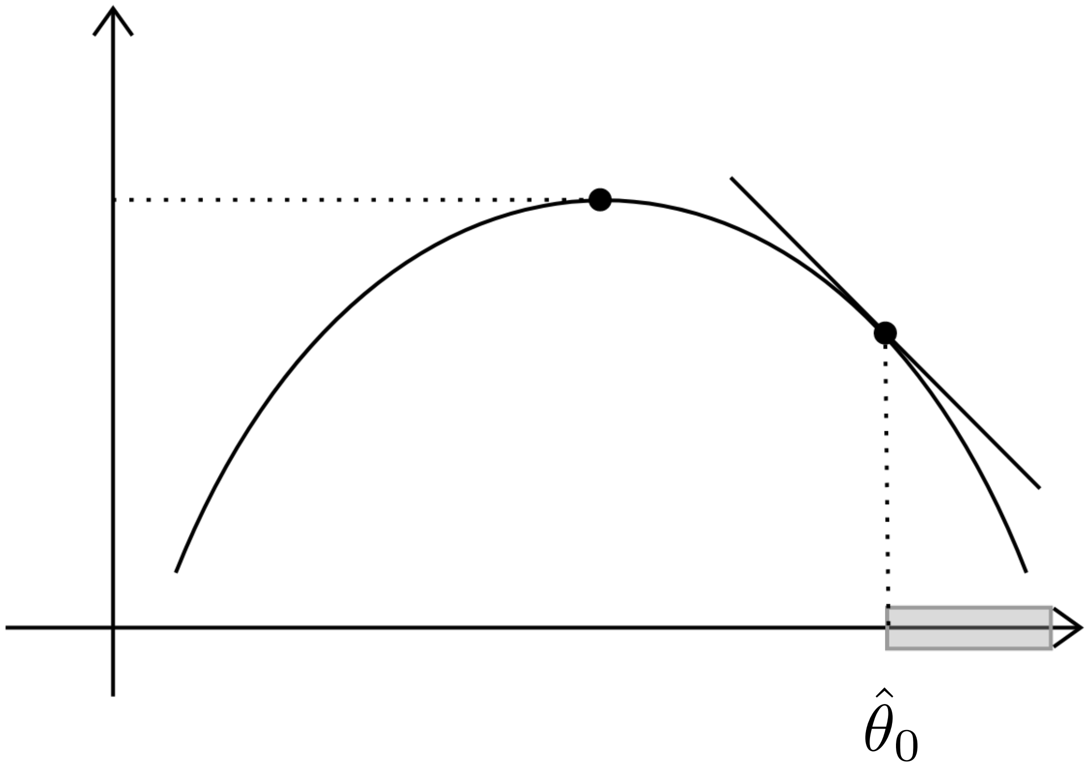
$$\mathcal{I}(\theta) = \mathop{\mathbb{E}}_{x|\theta} \left[\nabla^2 \mathcal{L}(x; \theta) | \theta \right] / T$$

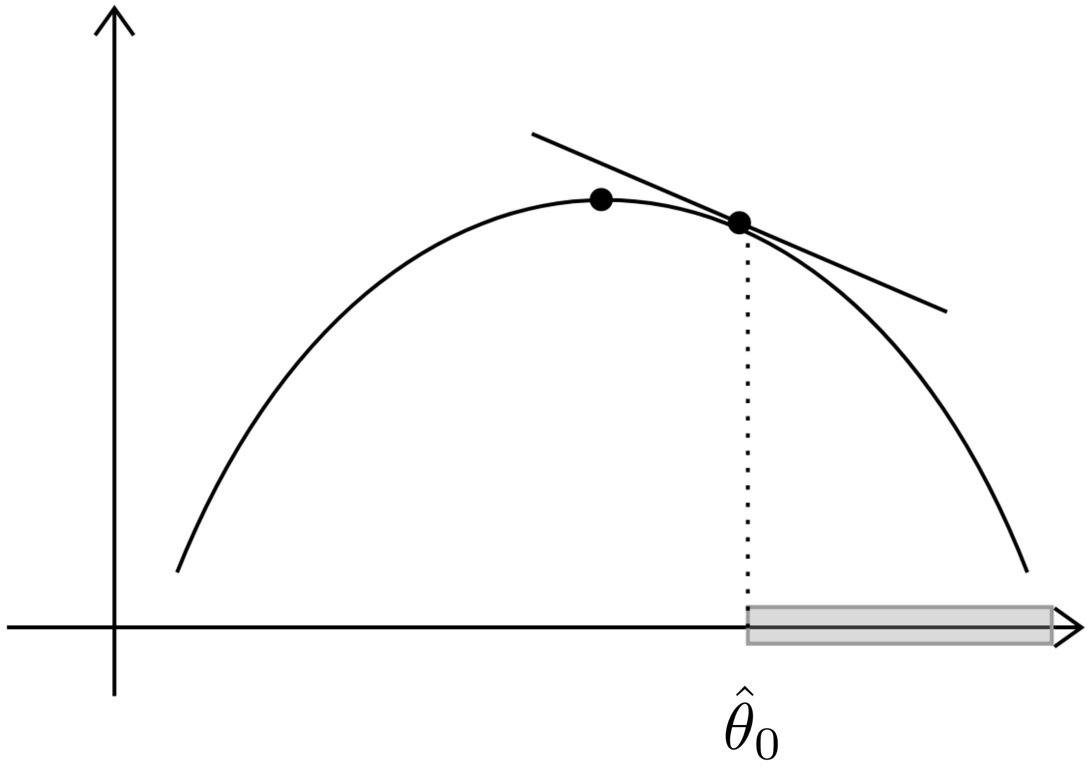
Wald's $\mathcal{X}^2$ Theorem (Informal)

Let $\hat{\theta}$ converge in distribution to a normal, and assume that the ML $\mathcal{I}(\hat{\theta})$ is a consistent estimator for $I(\hat{\theta})$. Under the null hypothesis, the Wald statistic will converge in distribution to a χ^2 distribution with degrees of freedom equal to the number of equality constraints.

Lagrange Multiplier Test

- Key idea: the Lagrange Multiplier associated with the constraint encodes how sensitive the likelihood is to its relaxation





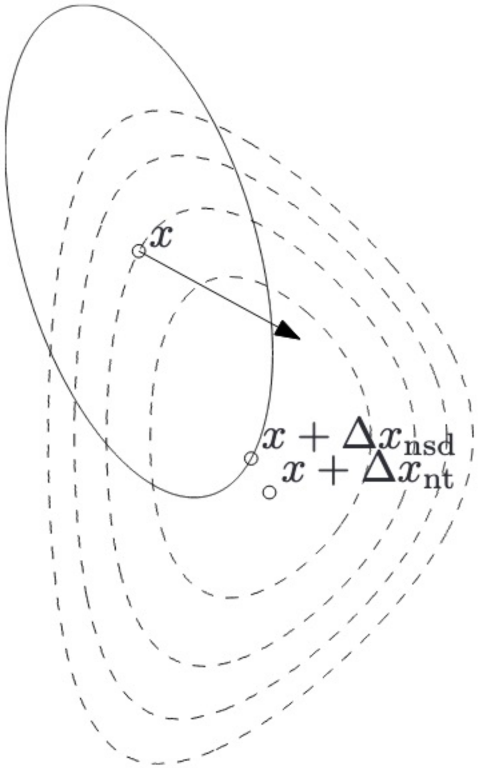
Some Optimization

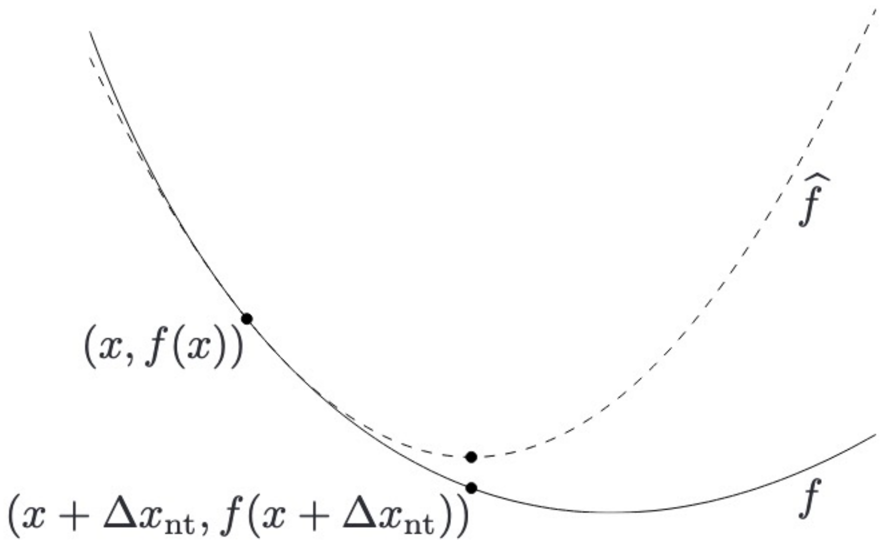
$$\text{Stationarity:} \qquad \nabla_x \underbrace{\ell(\hat{\theta}, \lambda^*, \nu^*)}_{\text{Lagragian}} = 0$$

$$\ell(\theta, \lambda, \nu) = \mathcal{L}(\theta; y) + \cancel{\lambda^\mathsf{T} g(x)} - \nu^\mathsf{T} h(x)$$

$$\underbrace{\nu^{*\mathsf{T}} \nabla_\theta h(\hat{\theta}_0)}_{\substack{\text{Score Function} \\ s(\hat{\theta}_0)}} = \nabla_\theta \mathcal{L}(\hat{\theta}_0)$$

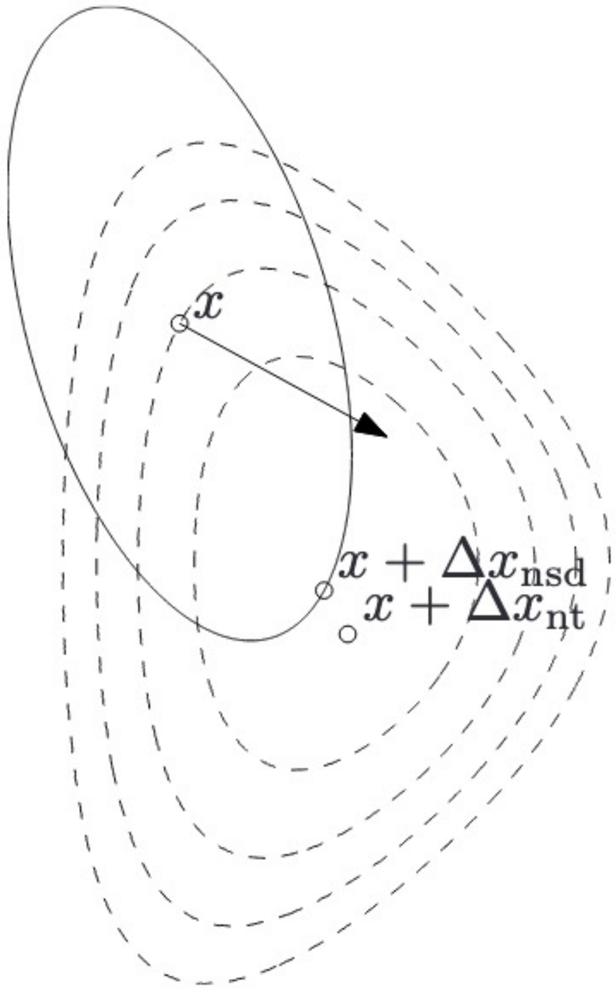
Newton Steps





Figures from: Convex Optimization: Stephen Boyd and Lieven Vandenberghe. Cambridge University Press 2004

Newton Steps



$$\Delta\theta_{nt} = -(\nabla^2_\theta \mathcal{L}(\hat{\theta}_0))^{-1} \nabla_\theta \mathcal{L}(\hat{\theta}_0)$$

$$\mathcal{L}(\hat{\theta}; y) - \mathcal{L}(\hat{\theta}_0; y) \approx \frac{1}{2} \underbrace{\nabla_\theta \mathcal{L}(\theta_0^*)^\mathsf{T} (\nabla^2_\theta \mathcal{L}(\theta_0^*))^{-1} \nabla_\theta \mathcal{L}(\theta_0^*)}_{\substack{\text{Newton Decrement} \\ \text{(Best 2nd Order Taylor Estimate)}}}$$

Approximation gets better as you $\hat{\theta_0}$ gets closer to $\hat{\theta}$, and is invariant to changes in coordinate system.

Figure from: Convex Optimization: Stephen Boyd and Lieven Vandenberghe. Cambridge University Press 2004

Back to Lagrange Multiplier test

$$\begin{aligned}\mathrm{LM} &= s(\theta_0^*)^\mathsf{T}\mathcal{I}(\theta_0^*)s(\theta_0^*) = \|\Delta\theta_{nt}\|_2 \\ &= \nu^{*\mathsf{T}}\nabla_\theta h(\theta_0^*)\left[\nabla_\theta^2(\mathcal{L}(\theta_0^*))\right]^{-1}(\theta_0^*)\nabla_\theta^\mathsf{T} h(\theta_0^*)\nu^*\end{aligned}$$

Rao's $\mathcal{X}^2$ Theorem (Informal)

Let θ^* converge in distribution to a normal, and assume that the ML $\mathcal{I}(\theta^*)$ is a consistent estimator for $I(\theta)$. Under the null hypothesis, the lagrange multiplier statistic will converge in distribution to a χ^2 distribution with degrees of freedom equal to the number of equality constraints.

Finite sample inequality

$$\mathrm{LM} \quad \leqslant \quad \mathrm{LR} \quad \leqslant \quad \mathrm{W}$$

- The “right” one to use depends on which optimization problem is easiest to compute
  - If both are easy, the likelihood ratio test should be preferred
  - If the restricted version is easier, then the lagrange multiplier test should be preferred
  - If The unrestricted version is easy, then the Wald test should be preferred.

Generalized Likelihood Ratio Tests

- The nested property of the model greatly simplifies analysis, but it is unknown when this condition can be relaxed and limiting distributions still resemble chi square distributions

Likelihood Ratios for Out-of-Distribution Detection

Jie Ren[*,†]
Google Research
jjren@google.com

Peter J. Liu[‡]
Google Research
peterjliu@google.com

Emily Fertig[†]
Google Research
emilyaf@google.com

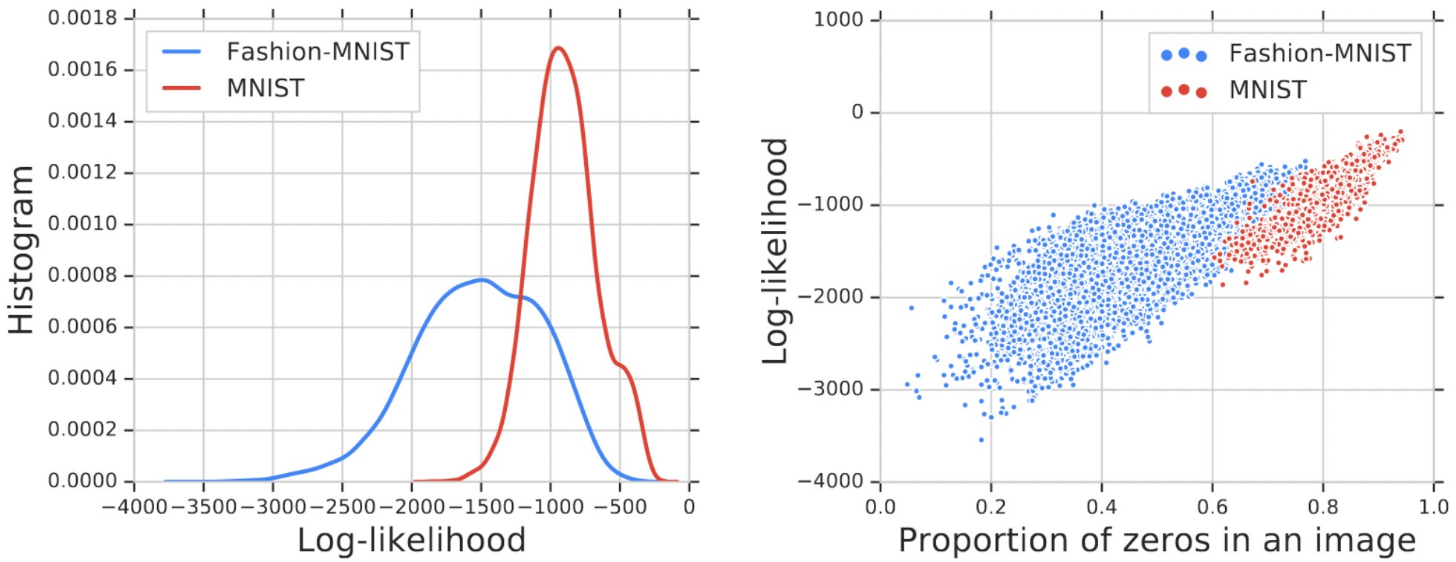
Jasper Snoek
Google Research

Ryan Poplin
Google Research

Mark A. DePristo
Google Research

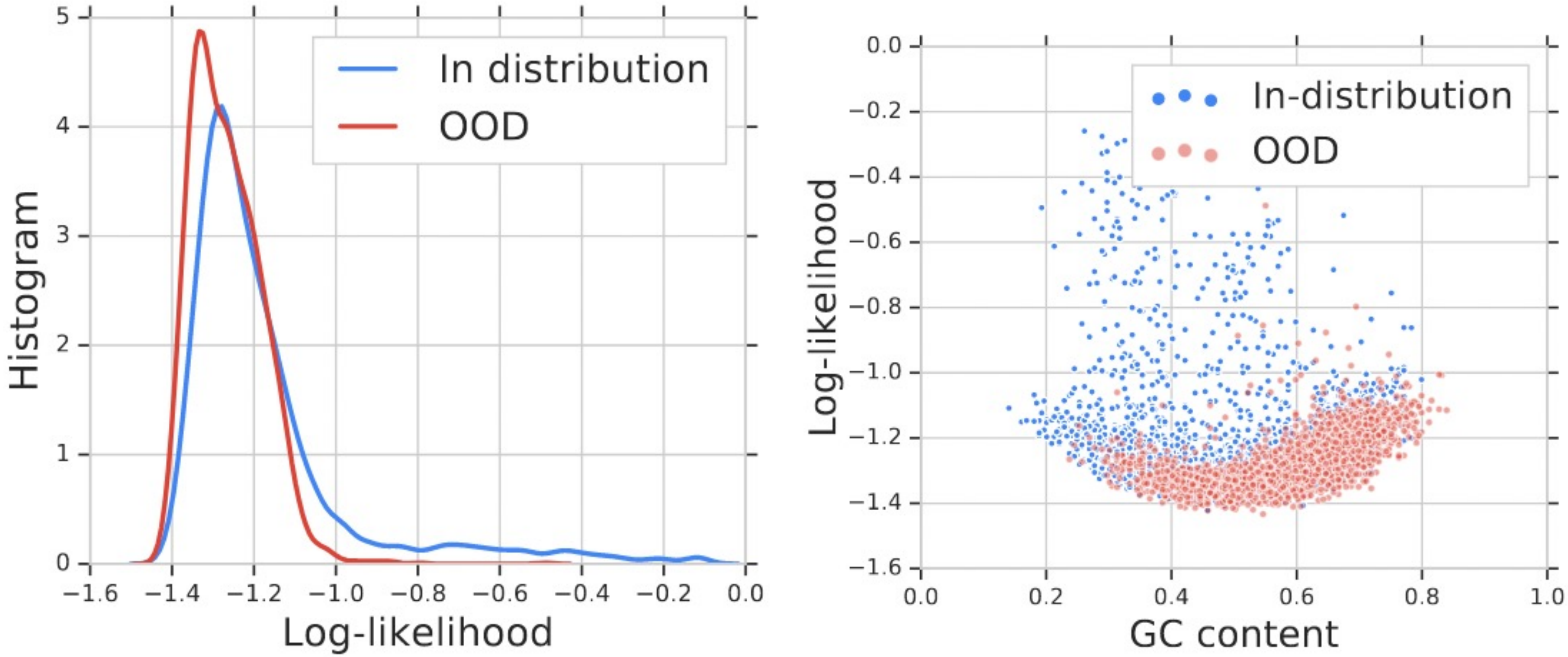
Goals and Problem Setup

- Does the input we’re evaluating on even come from the same distribution training dataset?
  - Oftentimes the likelihoods overlap enough that we cannot simply use likelihood alone



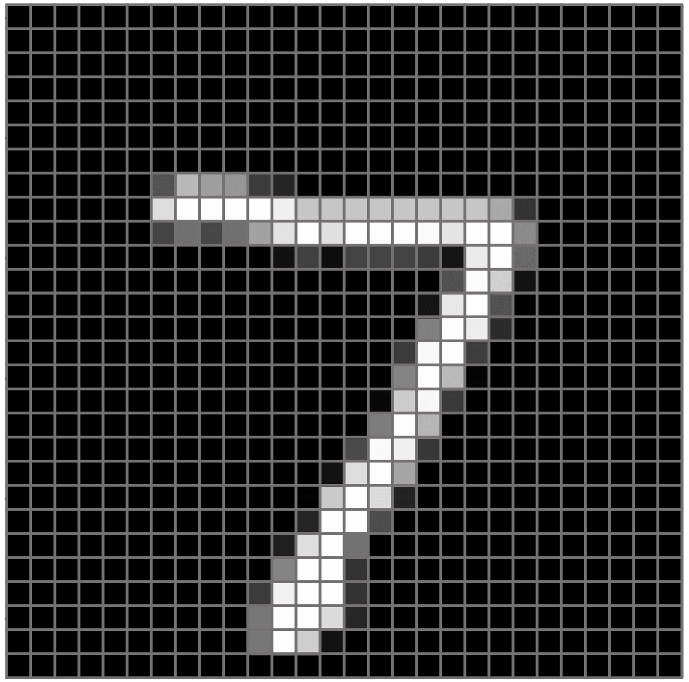
Each dot on right image in a single sequence input whose log-likelihood is evaluated and plotted on the histogram on the left

Dramatically worse example



- Situation is worse when data likelihoods overlap dramatically, as it does in DNA sequences

Problem Setup



Data label does not matter here!

- In distribution data is assumed to be generated from mixture model with latent states: **b**ackground, **s**emantic
  - Each MNIST image pixel either comes from a background distribution or a semantic distribution
  - The likelihood of observing a particular image is equal to a product

$$p(\mathbf{x}) = p(\mathbf{x_B})p(\mathbf{x_S})$$

- This assumption makes as much sense as the letters chosen

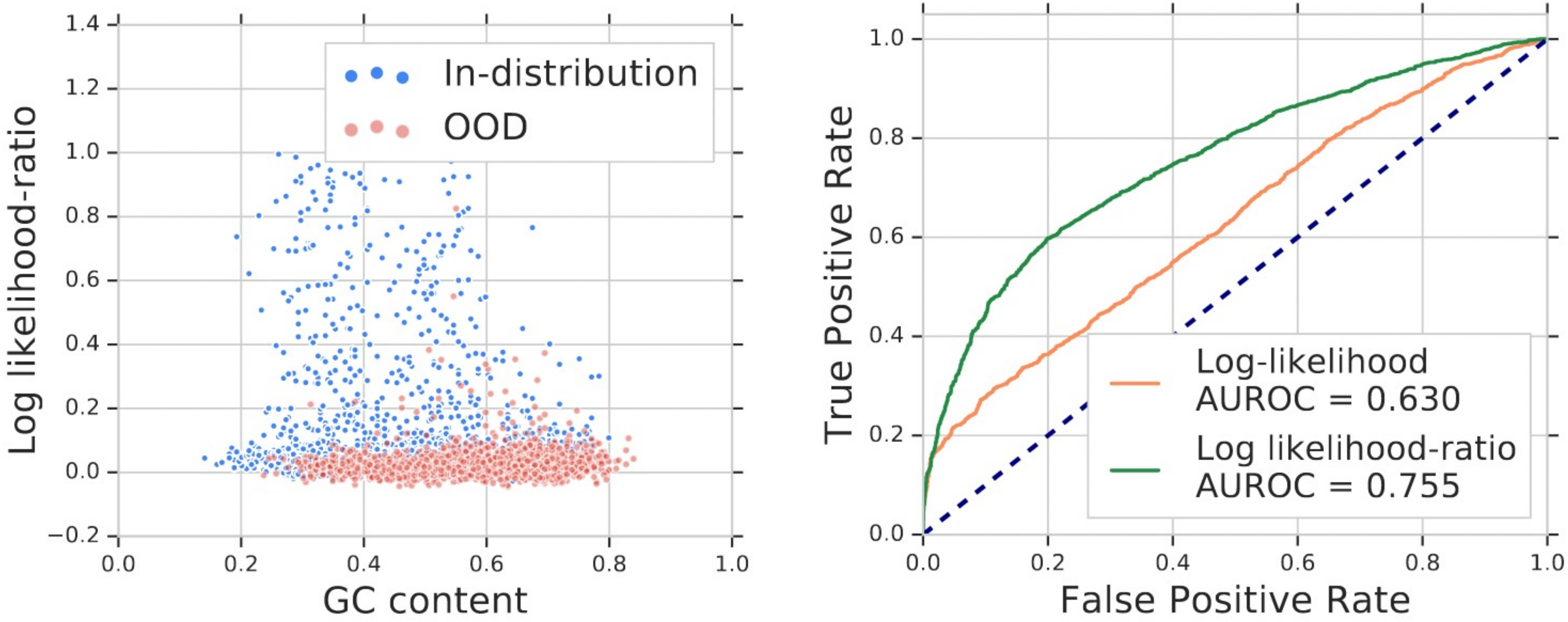
Assume we accept the independence assumption…

$$\text{“LR”} \quad = \quad \log\left(\frac{p_\theta(\mathbf{x})}{p_{\theta_{\text{“ “}0\text{” ”}}}(\mathbf{x})}\right) = \log\left(\frac{p_\theta(\mathbf{x_B})p_\theta(\mathbf{x_S})}{p_{\theta_0}(\mathbf{x_B})p_{\theta_0}(\mathbf{x_S})}\right)$$

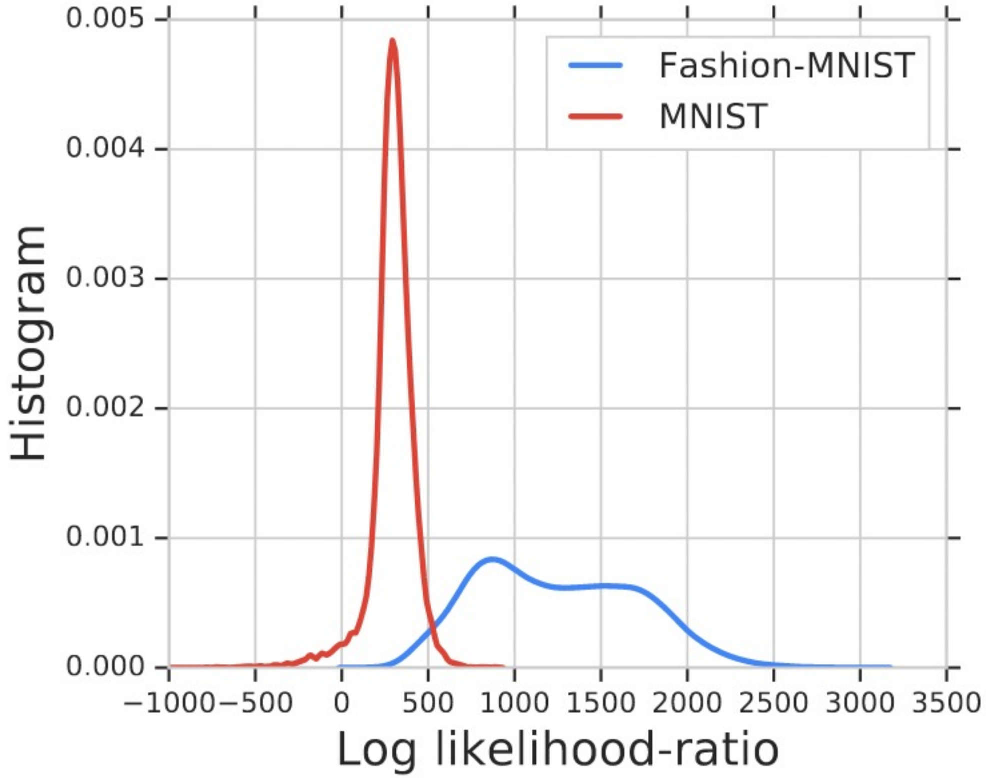
- A likelihood ratio is defined between two models
  - One trained normally
  - The other trained using the data plus some noise. This noise is supposed to help the second model capture “general background statistics”
    - These general background statistics somehow only barely affect the background term associated with the noise trained model so…

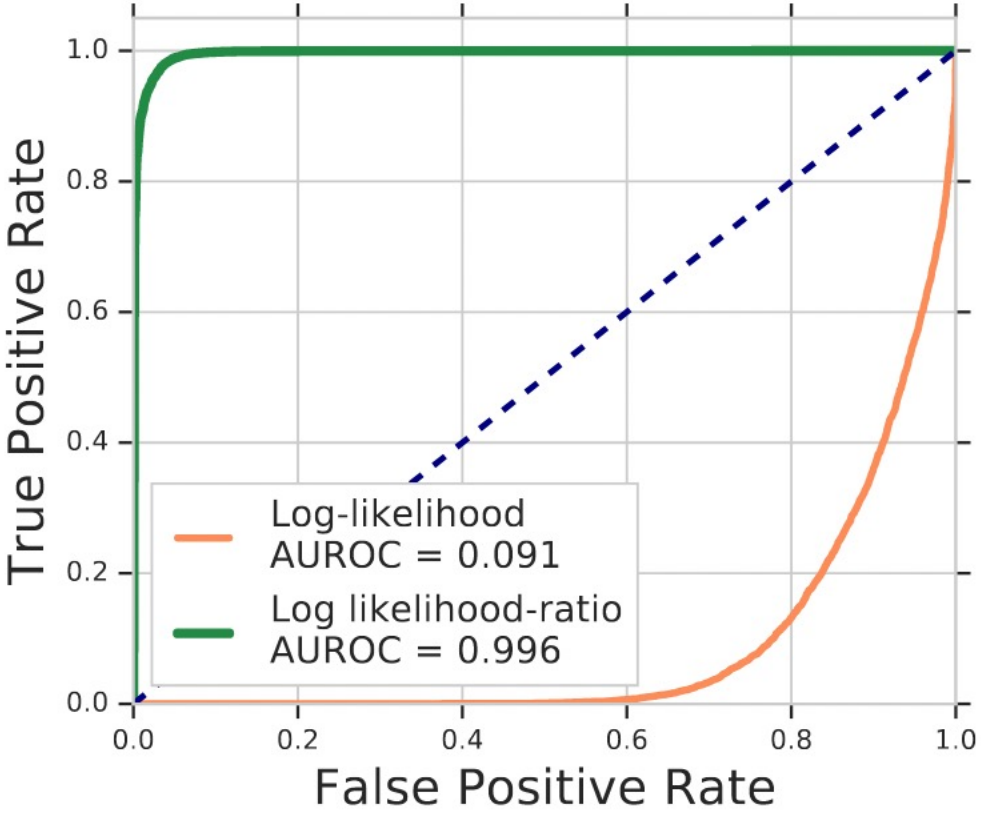
$$\mathrm{LR} = \log\left(\frac{p_\theta(\mathbf{x})}{p_{\theta_0}(\mathbf{x})}\right) \approx \log\left(\frac{p_\theta(\mathbf{x_s}))}{p_{\theta_0}(\mathbf{x_s})}\right)$$

Experimental Results for OOD DNA detection

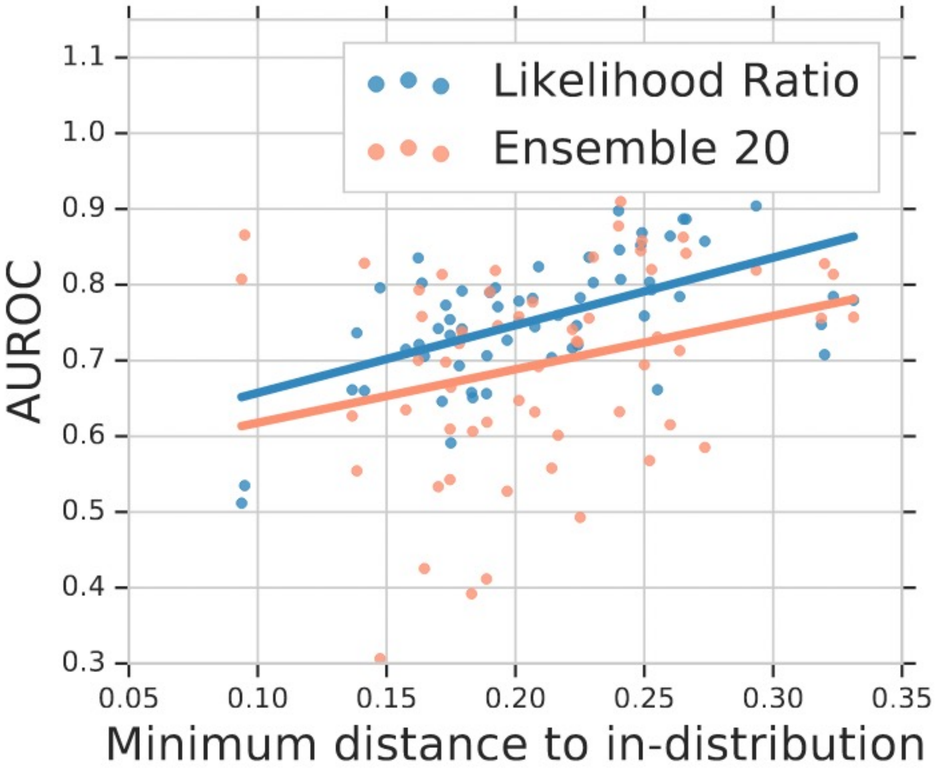


Non-symmetry of train/evaluation set





Experimental Results for OOD DNA detection



Conclusions…

- Works better than state of the art
- The extremely strong background/semantic assumption seems almost reasonable in context of state of art

- HUGE separation between theory and practice

Clustering Using Likelihood Ratios

- Paper predates k-means and expectation maximization (uses many of the same ideas!), but ideas from it are very elementary and have strong geometric interpretation

CLUSTERING METHODS BASED ON LIKELIHOOD RATIO CRITERIA

A. J. SCOTT[1] AND M. J. SYMONS

Department of Biostatistics, University of North Carolina, Chapel Hill, N. C. 27514, U. S. A.

Algorithm Sketch

- Objective:

$$\underset{C_g, \bar{y}_g}{\text{minimize}} \quad \sum_{g=1}^{G} \sum_{i \in C_g} (y_i - \bar{y}_g)^{\mathsf{T}} \Sigma_g^{-1} (y_i - \bar{y}_g)$$

Minimize the sum of distances to the center of each cluster, measured under a Mahalanobis distance specified by a known covariance

Algorithm Sketch

Before the time of EM/K-means, but ideas are practically the same:

- The cluster label assigned to a point is the cluster it is closest to (under Mahalanobis distance, not under Euclidean distance)
- An estimate of the covariance associated with each cluster can be formed via the sample covariance of the points assigned to each cluster
    - Simpler spherical estimation of each covariance matrix follows from using the frobenius norm
- This was before the time of easy computation, so each “iteration” is a combination of visual analysis, computer, and other heuristics developed at the time

Cluster analysis

- Works about as well as any other modern algorithm, but predates the era of cheap computation
- Major points of discussion involve decision boundaries and the heuristics used to find them