Are you using test log-likelihood correctly?

Sameer K. Deshpande* (Wisconsin) Soumya Ghosh* (IBM Research)
Tin D. Nguyen* (MIT) Tamara Broderick (MIT) **Equal contribution*

Overview

Test log-likelihood has been used to

- Compare different approximate inference algorithms
- Compare different predictive models

Our Contribution

Examples demonstrating comparisons based on test log-likelihood can *contradict* comparisons according to other objectives

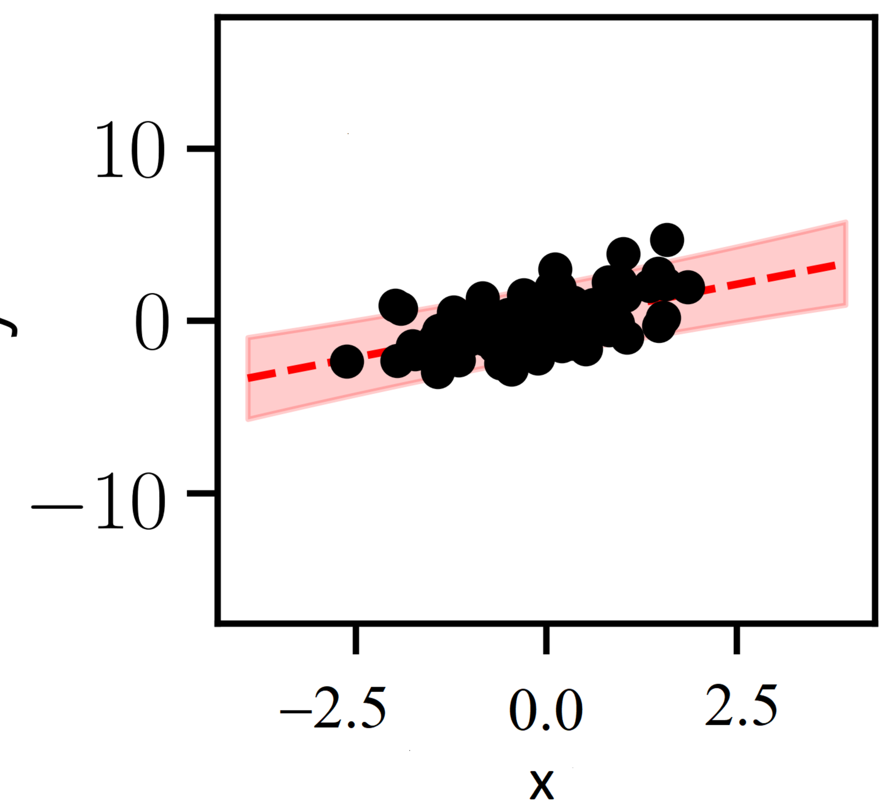
Example: Higher test log-likelihood, different inferential conclusion

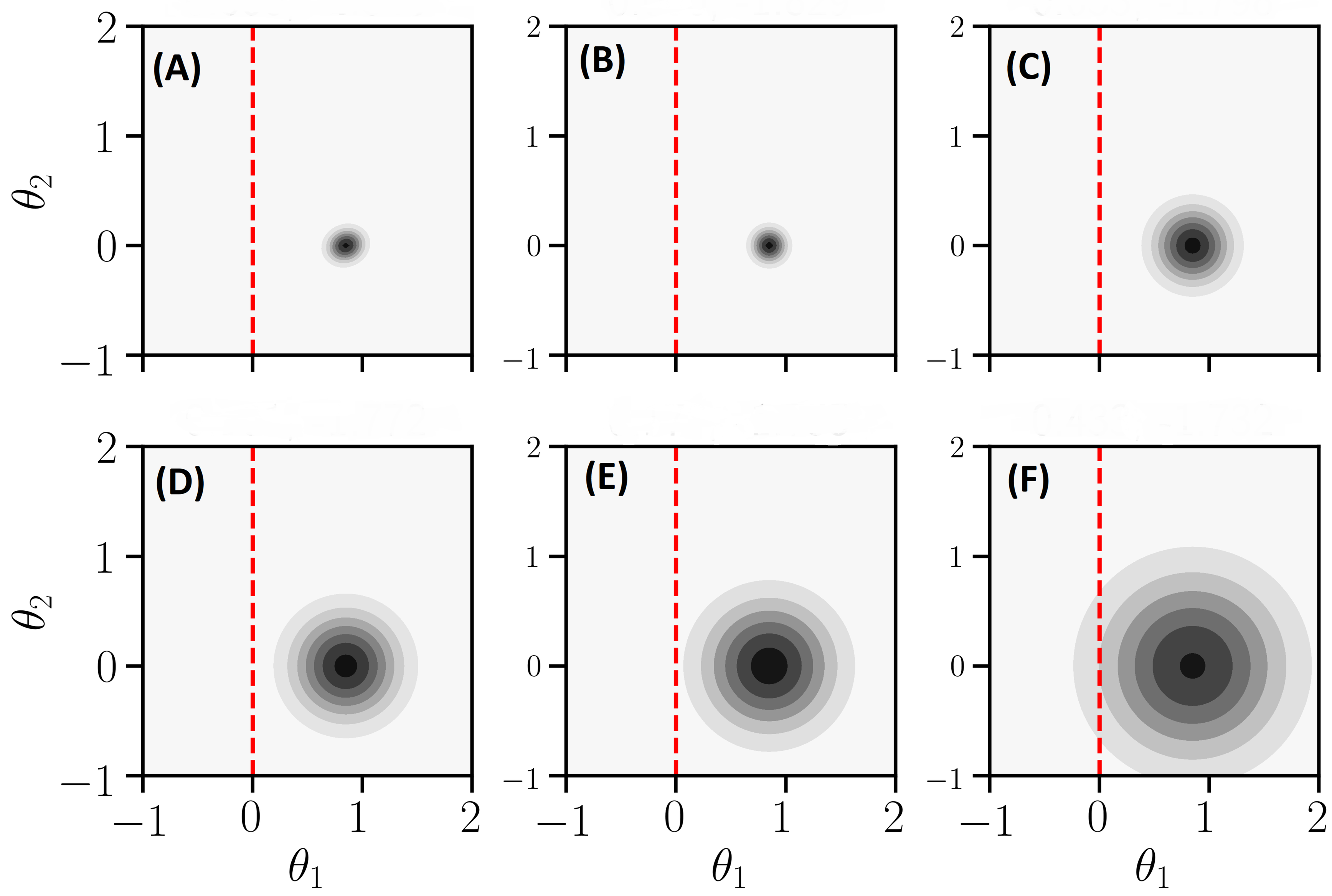
Data generating process: $x_n \sim \mathcal{N}(0,1), \quad y_n|x_n \sim \mathcal{N}(x_n, \log(1+\exp(x_n)))$ for $n = 1, \ldots, 100$

Model: $\theta \sim \mathcal{N}(\mathbf{0}, I_2) \quad y_n|\theta \sim \mathcal{N}(\theta_2 + \theta_1 x_n, 1)$. *In practice, all models are mis-specified.*

Under the posterior $\theta|\{(x_n,y_n)\}_{n=1}^{100}$, the 95% credible interval for θ_1 excludes 0 — see panel A below.

We say that a posterior approximation is *good* if it makes the same decision as the exact posterior (in this case, the 95% credible interval for the approximation also excludes 0).





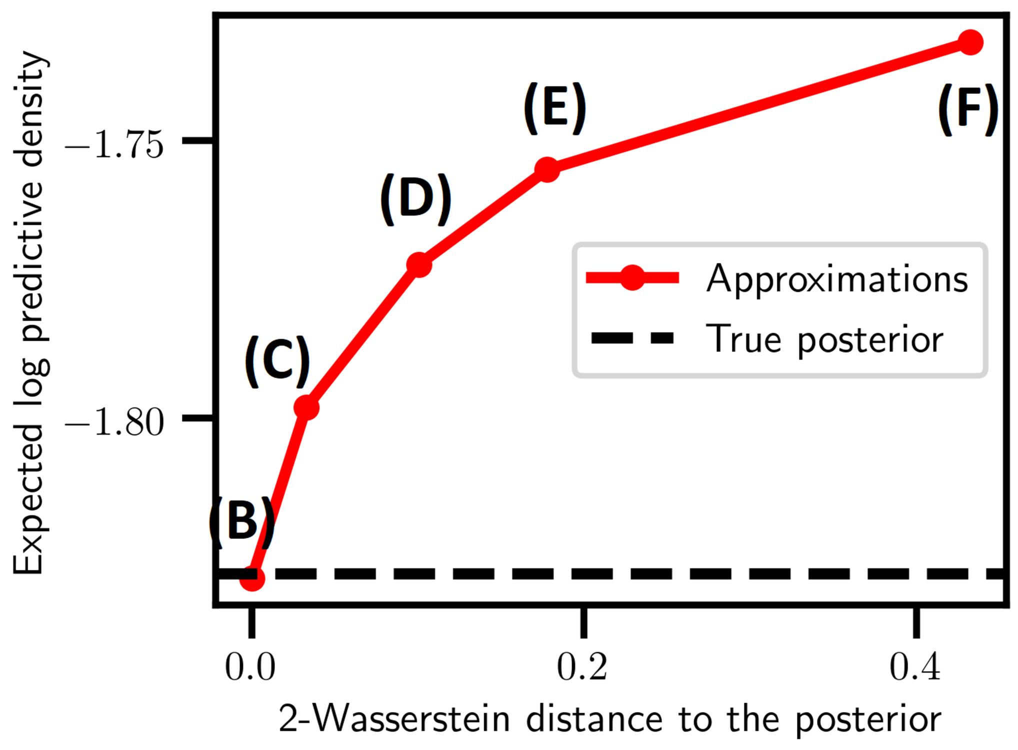
The credible interval under approximations A through E excludes 0.

The interval under approximation F includes 0.

C is better than F in terms of matching the decision under the posterior.

Moving from B through F increases the test log-likelihood.

F is better than C in terms of test log-likelihood.



y-axis measures the test log-likelihood evaluated on 10,000 test points.

x-axis measures the 2-Wasserstein distance to the posterior.

Find out more!

The paper also shows that comparison between different predictive models based on test log-likelihood can contradict the comparison based on squared errors.

Contact us at sameer.deshpande@wisc.edu, ghoshso@us.ibm.com, tdn@mit.edu, tamarab@mit.edu.