Statistical Model Criticism using Kernel Two-Sample Tests

Lloyd, Ghahramani

Arthur Gretton's notes, with bonus content

January 15, 2016

What the paper is about

- A method for discovering how well a model fits data.
- Uses MMD, including witness function to visualise differences in data generated from model and observed data
- Applied to:
  - Samples of digits generated from RBMs and DBNs
  - Evaluating the performance of the automated statistician
- Bonus content (not from paper) : testing goodness of fit of a model without drawing samples from it.

P-values

Given data $Y^{obs} := \left(y_i^{obs}\right)_{i=1}^{n}$. Model is M with parameters θ, construct statistic T of the data, "whose distribution does not depend on θ." (difficult! Example next slide)

There are several plausible definitions for p-values.

$$p_{freq}(Y^{obs}) = P(T(Y) \geq T(Y^{obs})) \qquad Y \sim p(Y|\theta, M) \quad \text{for any } \theta.$$

Prior predictive p-value:

$$Y \sim \int p(Y|\theta, M)p(\theta|M)d\theta.$$

Posterior predictive p-value:

$$Y \sim \int p(Y|\theta, M)p(\theta|Y^{obs}, M)d\theta.$$

Plug-in p-value:

$$Y \sim \int p(Y|\hat{\theta}, M) \qquad \hat{\theta} = \arg\max p(\theta|Y^{obs}, M).$$

Last two: how surprising is the data Y^{obs} even after you've seen it?

Aside: Choice of the statistic T

One example from the paper of Rubin, "Bayesianly justifiable and relevant frequency calculations for the applied statistician", p. 1168:

- Model is Gaussian
- Truth is Cauchy distribution
- A statistic is:

$$T(X) = \frac{X_{[10]} - X_{[9]}}{X_{[9]} - X_{[8]}}$$

(ratio of gaps between order statistics).

Idea for this paper

Idea for this paper: Fit the model to the observations, draw samples from the model (...wasteful?), compare samples with observations.

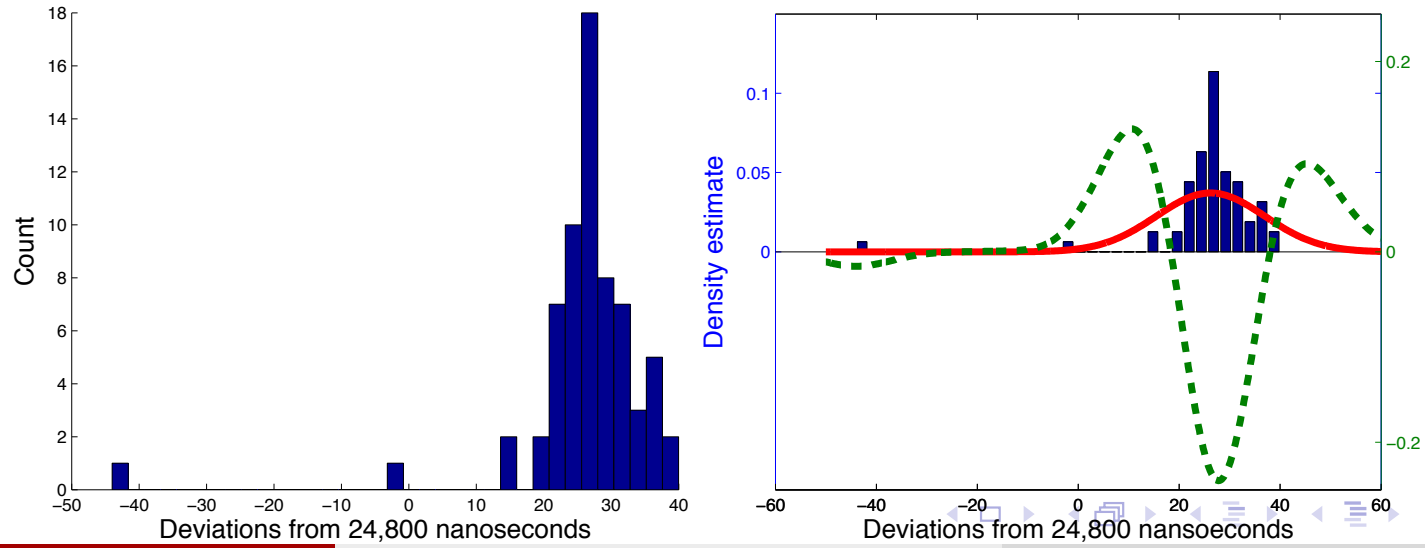
Reminder: MMD witness function

The maximum mean discrepancy is:

$$\mathrm{MMD}(P, Q) := \sup_{\|f\| \leq 1} E_P f - E_Q f.$$

When samples $\{x_i\}_{i=1}^n \sim P$ and $\{y_i\}_{i=1}^m \sim Q$, then an estimate of the witness function is:

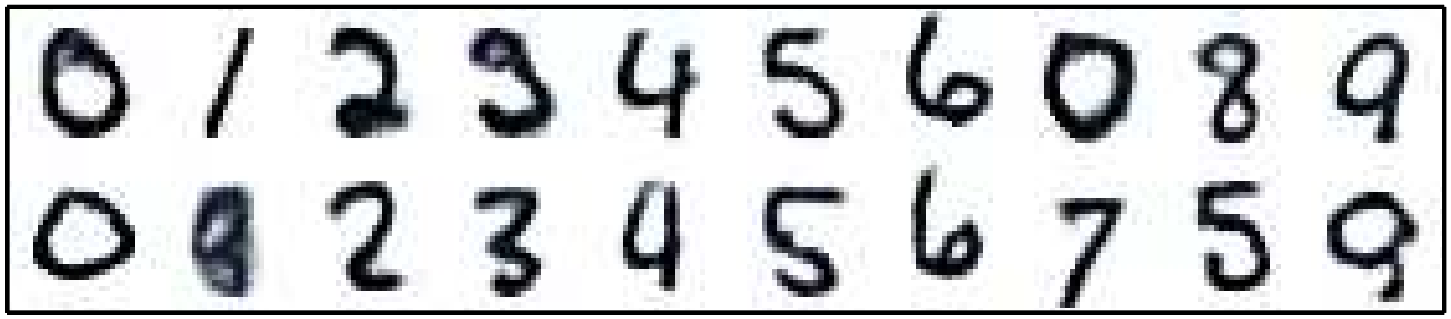
$$\hat{f} \propto \sum_{i=1}^{n} k(x_i, \cdot) - \sum_{j=1}^{m} k(y_i, \cdot).$$



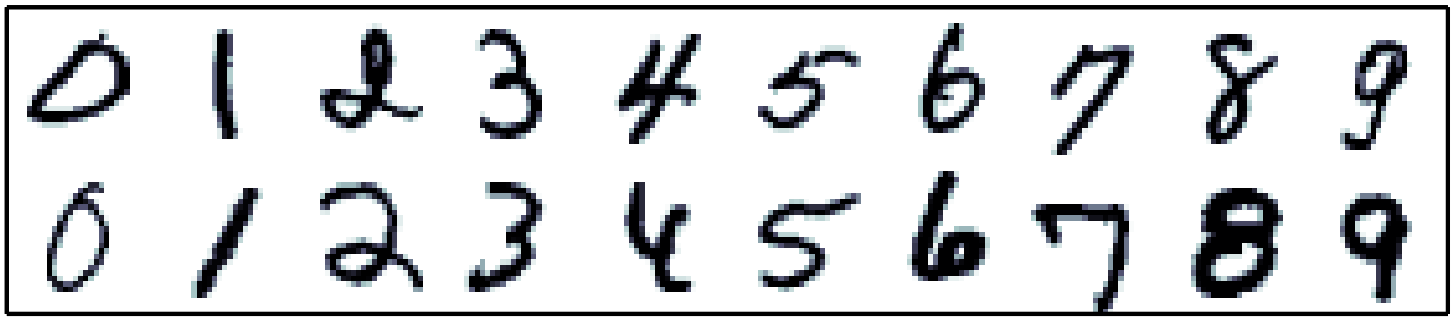
Finding poor sample regions, restricted Boltzmann machines

High and low values of witness function for RBM with 500 hidden units (samples aggregated from 1500 RBNs, each digit generated independently from clamping label)

High witness function, representative samples:

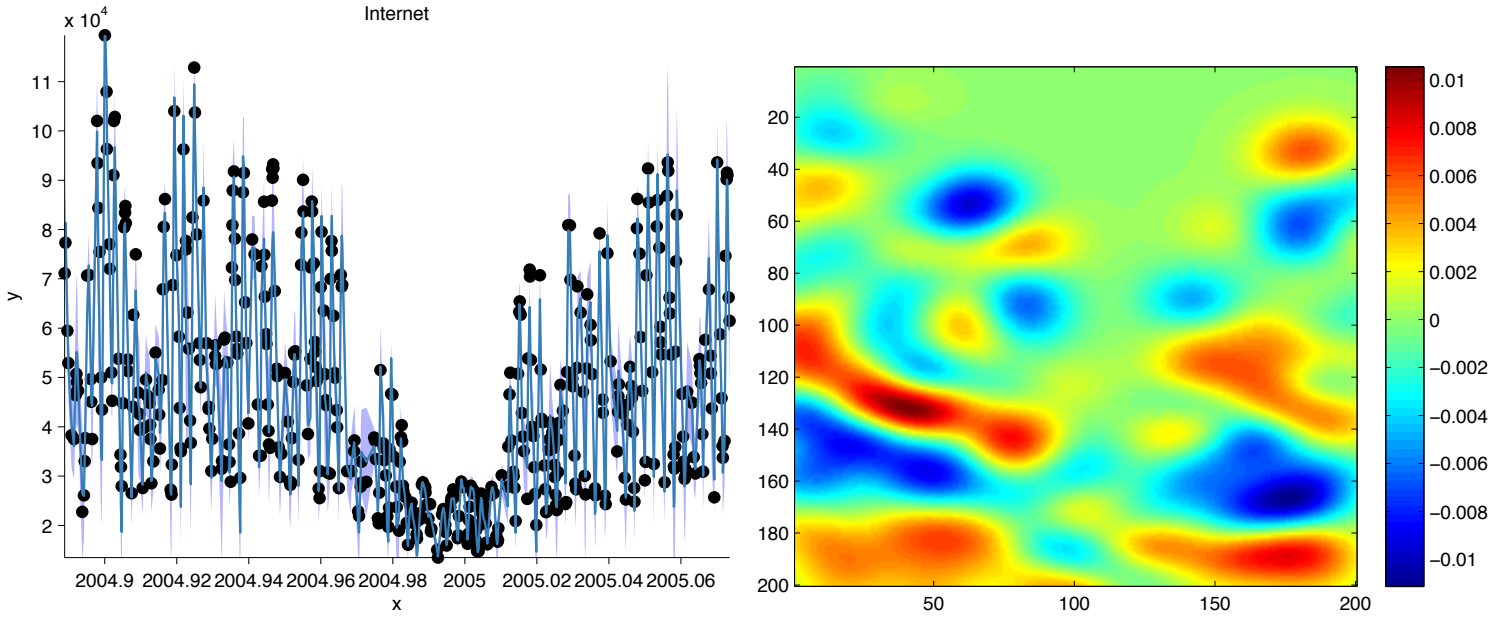


Low witness function, representative samples:



Finding poor sample regions, automated statistician

Gaussian process regression. Here Gaussian noise model is violated. Witness function has high amplitude where the Gaussian noise model does not match the samples.



Bonus content: goodness of fit without sample from model?

Define the Stein operator $T_q : \mathcal{F} \to \mathbb{R}^d$ such that

$$T_q f = \nabla \log q(x) f(x) + \nabla f(x).$$

Suppose $q(x)f(x) \to 0$ at the boundaries of $\mathcal{X}$. Then $\mathbb{E}_q T_q f = 0_d$, the $d \times 1$ vector of zeros.

Bonus content: goodness of fit without sample from model?

Define the Stein operator $T_q : \mathcal{F} \to \mathbb{R}^d$ such that

$$T_q f = \nabla \log q(x) f(x) + \nabla f(x).$$

Suppose $q(x)f(x) \to 0$ at the boundaries of $\mathcal{X}$. Then $\mathbb{E}_q T_q f = 0_d$, the $d \times 1$ vector of zeros.

Proof: integration by parts:

$$\begin{aligned}
\mathbb{E}_q T_q f &= \int \left[\nabla \left(\log q(x)\right) f(x) + \nabla f(x)\right] q(x) dx \\
&= \int \left[\frac{\nabla q(x)}{q(x)} f(x) + \nabla f(x)\right] q(x) dx \\
&= \int \left[\nabla q(x) f(x) + \nabla f(x) q(x)\right] dx \\
&= \left[\ldots [q(x)f(x)]_{-B_i}^{B_i} \ldots\right] + \int \left[-\nabla f(x) q(x) + \nabla f(x) q(x)\right] dx \\
&= 0_d.
\end{aligned}$$

Bonus content: MMD using Stein operator

MMD with function class restricted via the Stein operator:

$$d_q(p) = \sup_{\|f\| \leq 1} \left(\mathbb{E}_p T_q f - \mathbb{E}_q T_q f \right) = \sup_{\|f\| \leq 1} \mathbb{E}_p T_q f$$

Bonus content: MMD using Stein operator

MMD with function class restricted via the Stein operator:

$$
\begin{aligned}
d_q(p) &= \sup_{\|f\|\leq 1} \left(\mathbb{E}_p T_q f - \mathbb{E}_q T_q f\right) = \sup_{\|f\|\leq 1} \mathbb{E}_p T_q f \\
&= \sup_{\|f\|\leq 1} \mathbb{E}_p \left[\frac{d}{dx}\log q(x) f(x) + \frac{d}{dx} f(x)\right] \\
&= \sup_{\|f\|\leq 1} \left[\mathbb{E}_p \left\langle f, \left[\frac{d}{dx}\log q(x)\right] k(x,\cdot) + \frac{d}{dx} k(x,\cdot)\right\rangle\right] \\
&= \sup_{\|f\|\leq 1} \left\langle f, \mathbb{E}_p \left[\left[\frac{d}{dx}\log q(x)\right] k(x,\cdot) + \frac{d}{dx} k(x,\cdot)\right]\right\rangle \\
&= \|\xi\|
\end{aligned}
$$

where

$$
\xi := \int \left[\left(\frac{d}{dx}\log q(x)\right) k(x,\cdot) + \frac{d}{dx} k(x,\cdot)\right] p(x)dx \in \mathcal{F}
$$