Chapter 1

A method to determine where a statistical model is most wrong

This chapter. . .

1.1 Introduction

Statistical model checking or criticism[1]is an important part of a complete statistical analysis. When one fits a linear model to a data set via least squares a complete analysis includes computing e.g. Cook's distances **?** to identify influential points or plotting residuals against fitted values to identify non-linearity or heteroscedasticity. Similarly, modern approaches to Bayesian statistics view model criticism as in important component of an iterative cycle of model construction, inference and criticism (**?**).

As statistical models become more complex and diverse in response to the challenges of modern data sets there will be an increasing need for a greater range of model criticism procedures that are either automatic or generally applicable. This will be especially true as automatic modelling methods (e.g. **???**) and probabilistic programming (e.g. **???**) mature (1).

(1) What is the can
1997 reference?

Model criticism typically proceeds by choosing a statistic of interest, computing it on data and comparing this to a suitable null distribution. Ideally these statistics are chosen to assess the utility of the statistical model under consideration (see applied examples (e.g. **??**)) but this can require considerable expertise on the part of the modeller. We propose an alternative to this approach by using a statistic defined as a supremum over

[1] We follow Box (**?**) using the term 'model criticism' for similar reasons to O'Hagan (**?**).

a broad class of measures of discrepancy between two distributions, the maximum mean discrepancy (MMD) (e.g. ???). The advantage of this approach is that the discrepancy measure attaining the supremum automatically identifies regions of the data which are most poorly represented by the statistical model fit to the data.

We demonstrate this approach to model criticism on toy data sets, restricted Boltzmann machines[2] and deep belief networks[3] trained on MNIST digits and Gaussian process (e.g. ?) regression models trained on several time series. Our proposed method identifies discrepancies between the data and fitted models that would not be apparent from the predictive performance focused metrics one typically finds in a machine learning paper[4]. It is our belief[5] that more effort shoud be expended on attempting to falsify models fitted to data, using model criticism techniques or otherwise. Not only will this aid research in targeting areas for improvement but it should give greater confidence in any conclusions drawn from a model.

at is a canonical ref-
?

at is a canonical ref-
?

ould we be explicitly
ng to machine learn-

finding it difficult to
e right words here

1.2 Model criticism

Suppose we observe data $Y^{\text{obs}} = (y_i)_{i=1\ldots n}$ and we attempt to fit a model M with parameters θ. After performing a statistical analysis we will have either an estimate, $\hat{\theta}$, or an (approximate) posterior, $p(\theta \,|\, Y^{\text{obs}}, M)$, for the parameters. How can we check the validity of this analysis?

1.2.1 Criticising prior assumptions

The classical approach to model criticism is to attempt to falsify the null hypothesis that the data could have been generated by the model M for some value of the parameters θ i.e. $Y^{\text{obs}} \sim p(Y \,|\, \theta, M)$. This is typically achieved by constructing a statistic T of the data whose distribution does not depend on the parameters θ i.e. a pivotal quantity. The extent to which the observed data Y^{obs} differs from expectations under the model M could then be quantified with a tail-area based p-value

$$p_{\text{freq}}(Y^{\text{obs}}) = \mathbb{P}(T(Y) \geq T(Y^{\text{obs}})) \quad \text{where} \quad Y \sim p(Y \,|\, \theta, M) \quad \text{for any } \theta. \tag{1.1}$$

Analogous quantities in a Bayesian analysis are the prior predictive p-values of Box (?). The null hypothesis is replaced with the claim that the data could have been generated from the prior predictive distribution $Y^{\text{obs}} \sim \int p(Y \,|\, \theta, M) p(\theta \,|\, M) \mathrm{d}\theta$. A tail-area p-value can then be constructed for any statistic T of the data

$$p_{\text{prior}}(Y^{\text{obs}}) = \mathbb{P}(T(Y) \geq T(Y^{\text{obs}})) \quad \text{where} \quad Y \sim \int p(Y \,|\, \theta, M) p(\theta \,|\, M) \mathrm{d}\theta. \tag{1.2}$$

Both of these procedures construct a function of the data $p(Y^{\text{obs}})$ whose distribution under a suitable null hypothesis is uniform i.e. a p-value. The p-value quantifies how surprised one should be after observing data Y^{obs} having expected it to have been generated by the model. The different null hypotheses reflect the different uses of the word 'model' in frequentist and Bayesian analyses. A frequentist model is a class of probability distributions over data indexed by parameters whereas a Bayesian model is a joint probability distribution over data and parameters.

1.2.2 Criticising estimated models or posterior distributions

A constrasting method of Bayesian model criticism is the calculation of posterior predictive p-values (e.g. **??**) p_{post} where the prior predictive distribution is replaced with the posterior predictive distribution $Y \sim \int p(Y \,|\, \theta, M) p(\theta \,|\, Y^{\text{obs}}, M) \mathrm{d}\theta$. The corresponding test for an analysis resulting in a point estimate of the parameters $\hat{\theta}$ would use the plug-in predictive distribution $Y \sim p(Y \,|\, \hat{\theta}, M)$ to form the plug-in p-value p_{plug}.

These p-values quantify how surprised one should be if, after performing inference having observed data Y^{obs}, one were to observe new data whose value of the statistic T was equal to that of the original data. Put more simply, they quantify how surprising the data Y^{obs} is even after having observed it. A simple variant of this method of model criticism is to use held out data Y^*, generated from the same distribution as Y^{obs}, to compute a p-value i.e. $p(Y^*) = \mathbb{P}(T(Y) \geq T(Y^*))$. This quantifies how surprising the held out data is after having observed Y^{obs}.

1.2.3 Which type of model criticism should be used?

Different forms of model criticism are appropriate in different contexts, but we believe that posterior predictive and plug-in p-values will be most often useful for the types of statistical model considered in the machine learning literature. For example, suppose one is fitting a nonparametric or otherwise very flexible model to data e.g. a deep belief network. Classical p-values would assume a null hypothesis that the data could have been generated from some deep belief network. Since the space of all possible deep belief networks is very large it will be difficult to ever falsify this hypothesis. A more interesting null hypothesis to test in this example is whether or not our particular deep belief network can faithfully mimick the distribution it was trained on. This is the null hypothesis of posterior or plug-in p-values.

1.3 Model criticism for i.i.d. data using two sample tests

We assume that our data Y are i.i.d. samples from some unknown distribution $(y_i)_{i=1\dots n} \overset{\text{iid}}{\sim} p(y \mid \theta, M)$. After performing inference resulting in a point estimate of the parameters $\hat{\theta}$, the null hypothesis associated with a plug-in p-value is $(y_i^{\text{obs}})_{i=1\dots n} \overset{\text{iid}}{\sim} p(y \mid \hat{\theta}, M)$.

We can test this null hypothesis using a two sample test (e.g. **???**). In particular, we have samples of data $(y_i)_{i=1\dots n}$ and we can generate samples from the plug-in predictive distribution $(y_i^{\text{rep}})_{i=1\dots m} \overset{\text{iid}}{\sim} p(y \mid \hat{\theta}, M)$ and then test whether or not these samples could have been generated from the same distribution.

The methods for constructing p-values in the previous section all started with a statistic T of the entire data. This however is not necessary; a p-value is simply a random variable which has a uniform distribution under the null hypothesis. Instead of computing a statistic $T(Y^{\text{obs}})$ of the entire data we consider statistics of data points $t(y)$ and then consider the mean discrepancy between the two samples

$$\mathbb{E}(t(y^{\text{rep}})) - \mathbb{E}(t(y)). \tag{1.3}$$

This quantity measures how different the two distributions are on average as measured by the statistic t. The benefit of mean discrepancy measures is that we can analytically maximise the discrepancy over a large class of statistics t, allowing us to find the statistic that most shows any discrepancy between the two distributions.

1.4 Kernel maximum mean discrepancy (MMD) two sample tests

Consider the two sample problem. We are given samples $X = (x_i)_{i=1\dots m}$ and $Y = (y_i)_{i=1\dots n}$ drawn i.i.d. from distributions p and q respectively. Can we determine if $p \neq q$?

An answer to this problem is to consider maximum mean discrepancy (MMD) (**?**)

statistics (also called integral probability metrics (?))

$$\mathrm{MMD}(\mathcal{F}, p, q) = \sup_{f \in \mathcal{F}} (\mathbb{E}_{x \sim p}[f(x)] - \mathbb{E}_{y \sim q}[f(y)]) \tag{1.4}$$

where $\mathcal{F}$ is a set of functions. When $\mathcal{F}$ is a reproducing kernel Hilbert space (RKHS) the function attaining the supremum can be derived analytically and is called the witness function

$$f(x) = \mathbb{E}_{x' \sim p}[k(x, x')] - \mathbb{E}_{x' \sim q}[k(x, x')] \tag{1.5}$$

where k is the kernel of the RKHS.

Substituting this expression into equation (1.4) yields

$$\mathrm{MMD}^2(\mathcal{F}, p, q) = \mathbb{E}_{x,x' \sim p}[k(x, x')] + 2\mathbb{E}_{x \sim p, y \sim q}[k(x, y)] + \mathbb{E}_{y,y' \sim q}[k(y, y')]. \tag{1.6}$$

This expression only involves expectations of the kernel k which can be estimated empirically by

$$\mathrm{MMD}_b^2(\mathcal{F}, X, Y) = \frac{1}{m^2} \sum_{i,j=1}^{m} k(x_i, x_j) - \frac{2}{mn} \sum_{i,j=1}^{m,n} k(x_i, y_j) + \frac{1}{n^2} \sum_{i,j=1}^{n} k(y_i, y_j). \tag{1.7}$$

One can also estimate the witness function from finite samples

$$\hat{f}(x) = \frac{1}{m} \sum_{i=1}^{m} k(x, x_i) - \frac{1}{n} \sum_{i=1}^{n} k(x, y_i) \tag{1.8}$$

i.e. the empirical witness function is the difference of two kernel density estimates (e.g. ??). This means that we can interpret the witness function as showing where the estimated densities of p and q are most different.

1.4.1 Kernel choice

The nature of the two sample test defined by the kernel MMD depends on the choice of the kernel. In this paper we use the radial basis function kernel, also known as the squared exponential or exponentiated quadratic. This kernel encodes for smooth functions characterised by a typical lengthscale (e.g. ?). A typical heuristic for selecting the lengthscale is to use the median distance between all points as the lengthscale (e.g. ?). However, since we interpret the witness function as the difference of two kernel density estimates we also consider selecting the lengthscale which gives the best density estimates (see section 1.5.2).

1.4.2 Estimation of the null distribution

There are a number of different ways in which the null distribution of the MMD statistic (1.7) can be estimated (e.g. ?). We use the bootstrap variant for its simplicity and general applicability.

1.5 Examples on toy data

1.5.1 Newcomb's speed of light data

A histogram of Simon Newcomb's 66 measurements used to determine the speed of light (?) is shown on the left of figure 1.1. We consider fitting a normal distribution to this data by maximum likelihood.

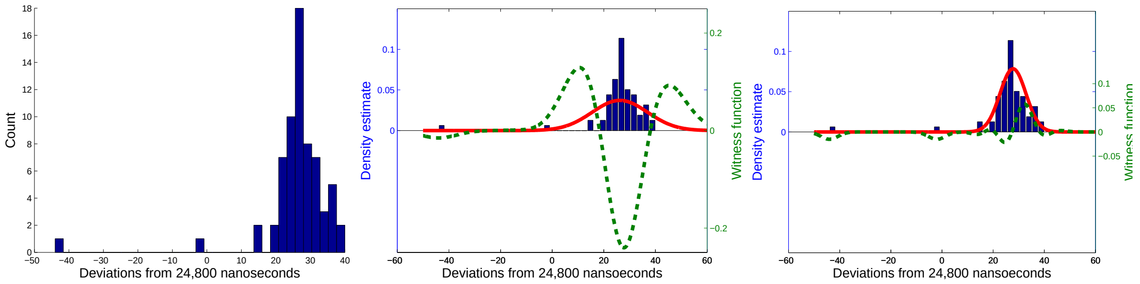

Figure 1.1: Left: Histogram of Simon Newcomb's speed of light measurements. Middle: Histogram together with density estimate (red solid line) and MMD witness function (green dashed line). Right: Histogram together with improved density estimate and witness function.

To perform a MMD two sample test we sampled 1000 points from the fitted distribution, used the median heuristic to select a lengthscale and estimated the null distribution using 1000 bootstrap replications. The estimated p-value of the test was less than 0.001 i.e. a clear disparity between the model and data. The data, fitted density estimate (the normal distribution) and witness function are shown in the middle of figure 1.1. The witness function has a trough at the centre of the data and peaks either side. This indicates that the fitted model has placed too little mass in its centre and too much mass outside its centre.

This suggests that we should modify our model by either using a distribution with heavy tails or explicitly modelling the possibility of outliers which could have resulted in the variance being over-estimated. However, to demonstrate some of the properties of the MMD statistic we make an unusual choice of fitting a Gaussian by maximum likelihood, but ignoring the two outliers in the data. The new fitted density estimate (the normal distribution) and witness function of an MMD test are shown on the right of figure 1.1. The estimated p-value associated with the MMD two sample test is roughly 0.5, despite the fitted model being a very poor explanation of the outliers. This demonstrates that the MMD test using a radial basis function kernel identifies dense discrepancies, rather than outliers. However, methods that are not robust to outliers (e.g. fitting a Gaussian by maximum likelihood) will likely show dense discrepancies that will be identified by the test.

1.5.2 High dimensional data

The interpretability of the witness functions comes from being equal to the difference of two kernel density estimates (1.8). In high dimensional spaces, kernel density estimation is a very high variance procedure that can result in poor density estimates (6) which will destroy the interpretability of the method. In response, we consider using dimensionality reduction techniques (7) before performing two sample tests. Note however that the statistical test derived from the MMD still has high power in high dimensions (**?**).

(6) what is a classi
tion?

(7) what is a classi
tion?

To test how the MMD statistic can be used for high dimensional data we generated synthetic data using the following recipe. 5 points in a 10 dimensional space were drawn at random from a random 4 dimensional subspace[2]. Data was generated as isotropic Gaussian distributions centred on 4 of the 5 points. Finally, data was centred on the fifth point drawn from an isotropic t-distribution with 2 degrees of freedom. In sum, the data is a mixture of Gaussians and a t-distribution.

We then fit a mixture of Gaussians (e.g. **?**) with 5 centres to the data and then generated samples from the fitted distribution in order to perform an MMD two sample test. We reduced the dimensionality of the data using principal component analysis (PCA), selecting the first two principal components. To ensure that the MMD test remains well calibrated we include the PCA dimensionality reduction within the bootstrap estimation of the null distribution. The data and posterior predictive samples are plotted on the left of figure 1.2. While we can see that one cluster is different from the rest, it is difficult to assess by eye if these distributions are different — due in part to the difficulty of plotting two sets of samples on top of each other.

[2] The details are not especially important; code for replication will be available upon publication

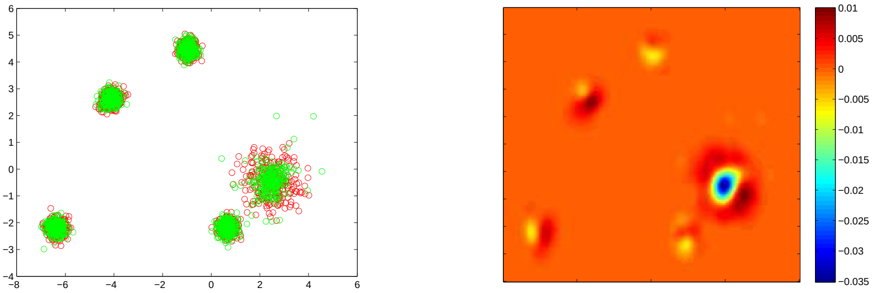

Figure 1.2: Left: PCA projection of synthetic high dimensional cluster data (green circles) and projection of samples from fitted model (red circles). Right: Witness function of MMD two sample test. The erroneously fit cluster is clearly identified.

Using the median heuristic to select a lengthscale results in a test that returns a p-value of 0.91, indicating that the test has not identified any discrepancies. Indeed, the lengthscale chosen by this heuristic is 4.4 which is of the order of the distances between clusters. The test therefore is blind to discrepancies smaller than distances between clusters[3], and since the mixture of Gaussians has correctly identified the 5 centres and sizes of the mixture distribution, the test does not find any discrepancies.

However, taking the density estimate interpretation of the witness function more seriously suggests choosing lengthscales that result in the best density estimates. We therefore selected a lengthscale by 5 fold cross validation using predictive likelihood of the kernel density estimate as the selection criterion. With this lengthscale the MMD test returns a p-value of 0.05 and the witness function (right of figure 1.2) clearly identifies the cluster that has been incorrectly modelled.

Presented with this discrepancy a statistical modeller might try a more flexible clustering model (e.g. **??**) (a mixture of t-distributions would work on this example). However, the p-value of the MMD statistic can also be made non-significant by fitting a mixture of 10 Gaussians. We mention this as a reminder that the test proposed here does not attempt to falsify a class of models, it tests only whether or not the data could plausibly have been generated by a particular fitted model.

[3] With enough data the test would eventually identify discrepancies on any scale — however, the required amount of data can easily be very large

1.6 Applications to real data and complex statistical models

1.6.1 What exactly do neural networks dream about?

"To recognize shapes, first learn to generate images" quoth Hinton (?). Restricted Boltzmann Machine (RBM) pretraining of neural networks was shown by ? to learn a deep belief network (DBN) for the data i.e. a generative model. In agreement with this observation, as well as computing estimates of marginal likelihoods and testing errors, it has been standard to demonstrate the effectiveness of a neural network by generating samples from the distribution it has learned.

When trained on the MNIST handwritten digit data, samples from RBMs and DBNs certainly look like digits, but it is hard to detect any systematic anomalies purely by visual inspection. We now use the kernel MMD two-sample test to investigate how faithfully RBMs and DBNs can capture the distribution over handwritten digits.

RBMs mistake the identity of digits

We trained an RBM with architecture $(784) \leftrightarrow (500) \leftrightarrow (10)$[4] using 15 epochs of persistent contrastive divergence PCD-15, a batch size of 20 and a learning rate of 0.1 (i.e. we used the same settings as the code available at the deep learning tutorial (?)). We generated 3000 independent samples from the learned generative model by initialising the network with a random training image and performing 1000 gibbs updates with the digit labels clamped[5] to generate each image (as in e.g. ?).

The top left of figure 1.3 shows twenty random samples[6] from this model. They certainly look mostly like digits, but has the true distribution over digits been faithfully captured? A priori the answer to this question is almost certainly no, but it is not immediately obvious how the learned distribution will deviate from the true distribution.

Since we generated digits from the class conditional distributions we compare each class separately. Rather than show plots of the witness function for each digit we summarise the witness function by examples of digits closest to the peaks and troughs of the

[4] That is, 784 input pixels and 10 indicators of the class label are connected to 500 hidden neurons.

[5] Without clamping the label neurons, the generative distribution is heavily biased towards certain digits.

[6] Specifically these are the activations of the pixel neurons before sampling sampling binary values. This is an attempt to be consistent with the grayscale input distribution of the images. Analogous discrepancies would be discovered if we had instead sampled binary pixel values.

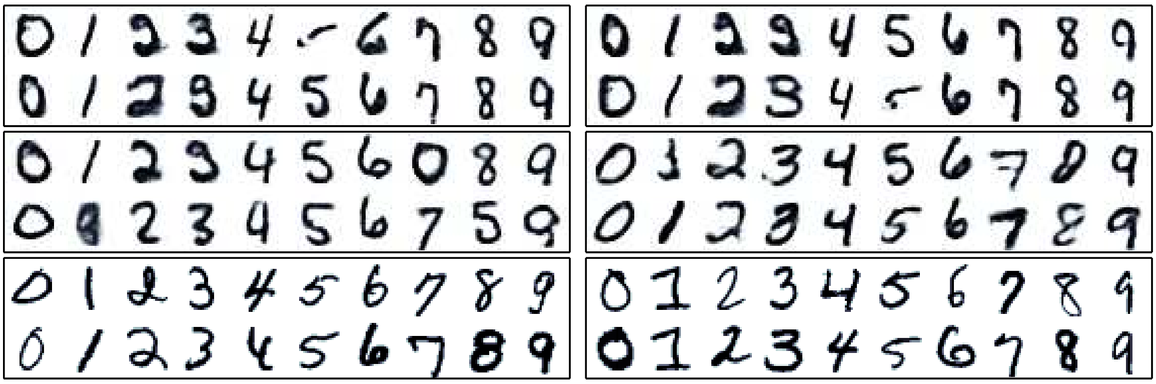

Figure 1.3: Top left: Random samples from an RBM. Top right: Troughs of the witness function for the RBM (digits that are over-represented by the model). Middle left: Troughs of the witness function for samples from 1500 RBMs. Middle right: Troughs of the witness function for the DBN. Bottom left: Peaks (digits that are under-represented by the model) of the witness function for samples from 1500 RBMs. Bottom right: Peaks of the witness function for the DBN.

witness function (the witness function estimate is differentiable so we can find the peaks and troughs by gradient based optimisation). We apply the MMD two-sample test to each class conditional distribution, using PCA to reduce to 2 dimensions and selecting the lengthscale via cross validation as in section 1.5.2.

The top right box of figure 1.3 shows the digits closest to the two most extreme troughs[7] of the witness function for each class; the troughs indicate where the fitted distribution over-represents the distribution of true digits. The estimated p-value for all tests was less than 0.001. The most obvious error with these digits is that the first 2 and 3 look quite similar.

To test that this was not just a poorly trained single RBM, we trained 1500 RBMs (with differently initialised pseudo random number generators) and generated one sample from each and performed the same tests. The estimated p-values were again all less than 0.001 and the summaries of the troughs of the witness function are shown in the middle left box of figure 1.3. On the first toy data example we observed that the MMD statistic does not highlight outliers and therefore we can conclude that RBMs are making consistent mistakes e.g. generating a 0 from the 7 distribution or a 5 when it should have been generating an 8.

[7] The exact ordering of the peaks and troughs is as follows. We partition the space by grouping samples where the witness function has the same sign and gradient based optimisation of the witness function starting from each sample would reach the same peak or trough. The contribution to the MMD from each of these groups is used to order the peaks and troughs.

DBNs have nightmares about ghosts

We now test the effectiveness of deep learning to represent the distribution of MNIST digits. In particular, we fit a DBN with architecture $(784) \leftarrow (500) \leftarrow (500) \leftrightarrow (2000) \leftrightarrow (10)$ using RBM pre-training and a generative fine tuning algorithm described in **?**. Performing the same tests with 3000 samples results in estimated p-values of less than 0.001 except for the digit 4 (0.150) and digit 7 (0.010). Summaries of the witness function troughs are shown in the middle right box of figure 1.3.

The witness function no longer shows any class label mistakes (except perhaps for the digit 1 which looks very peculiar) but the 2, 3, 7 and 8 appear 'ghosted' — the digits fade in and out. For comparison the bottom right box of figure 1.3 shows digits closest to the peaks of the witness function; there is no trace of ghosting.

Returning to the RBMs, we do not see ghosting either, but the digits nearest the witness function troughs are somewhat blurred (see bottom left box for comparison with peaks). Assuming that the top level associative memory of the DBN also suffers from blurring, this will result in occasionally incorrect neurons in the second hidden layer on the DBN. These incorrect bits will then propagate down the DBN resulting in spurious features in several visible neurons, resulting in ghosting.

Do we need to go deeper?

(!) Currently performing a p-values by depth experiment - unfortunately the parameters of the algorithms are very sensitive to depth so I need to make the experiments select these parameters ideally...

1.6.2 Testing non i.i.d. data

The test described so far applies when the model being tested has an i.i.d. predictive distribution. This is of course restrictive, so we now demonstrate how we can construct a test for a non i.i.d. model based on the MMD statistic. In particular we consider regression.

A test of local heteroscedasticity and non-normality for regression

We now assume that our data consists of pairs of inputs and outputs $(x_i^{\text{obs}}, y_i^{\text{obs}})_{i=1\ldots n}$. A typical formulation of the problem of regression is to estimate the conditional distribution of the outputs given the inputs $p(y \,|\, x, \theta)$. This is consistent with assuming that input-output pairs are generated i.i.d. from some distribution $p(y, x \mid \theta)$, but conditioned on observing the particular input values $(x_i^{\text{obs}})_{i=1\ldots n}$.

Following this observation, we might consider generating data from the plug-in conditional distribution $y_i^{\text{rep}} \sim p(y \mid x_i^{\text{obs}}, \hat{\theta})$ and computing the empirical MMD estimate (1.7) between $(x_i^{\text{obs}}, y_i^{\text{obs}})_{i=1\ldots n}$ and $(x_i^{\text{obs}}, y_i^{\text{rep}})_{i=1\ldots n}$. The only difference between this test and the MMD two sample test is that our data is generated from conditional distributions, rather than being i.i.d. . The null distribution of this statistic can be trivially estimated by sampling several sets of replicate data from the plug-in predictive distribution.

To demonstrate this test we apply it to 4 regression algorithms and 13 time series analysed in **?**. In this work the authors compare several methods for constructing Gaussian process (e.g. **?**) regression models. Example data sets are shown in figures 1.4 and 1.5. While it is clear that simple smoothing methods will fail to capture all of the structure in the data, it is not clear a priori how much better the more advanced methods will fair.

To construct p-values we use held out data using the same split of training and testing data as the interpolation experiment in **?**. Gaussian processes when applied to regression problems learn a joint distribution of all output values. However, this joint distribution information is rarely used; typically only the pointwise conditional distributions $p(y \mid x_i^{\text{obs}}, \hat{\theta})$ are used which is consistent with the test proposed here.

Table 1.1 shows a table of p-values for 13 data sets and 4 model construction methods. The four methods are Gaussian process regression using a squared exponential kernel (SE), trend-cyclical-irregular models (e.g. **?**) (TCI), spectral mixture kernels (**?**) (SP) and the method proposed in **?** (ABCD). Values in bold indicate a positive discovery after a Benjamini–Hochberg (**?**) procedure with a false discovery rate of 0.05 applied to each model construction method. SE, TCI and SP have a very similar pattern of significant p-values whereas ABCD has fewer significant p-values.

We now investigate the type of discrepancies found by this test by looking at the witness function (which can still be interpreted as the difference of kernel density estimates). Figure 1.4 shows the solar and gas production data sets, the posterior distribution of the SE fits to this data and the witness functions for the SE fit. The solar witness function has a clear narrow peak, indicating that the data is more dense than expected by the fitted model in this region. We can see that this has identified a region of low variability in the data i.e. it has identified local heteroscedasticity not captured by the model. Similar conclusions can be drawn about the gas production data and witness

Dataset	SE	TCI	SP	ABCD
Airline	0.36	**0.00**	0.07	0.15
Solar	**0.00**	**0.00**	**0.00**	0.05
Mauna	0.99	0.41	0.34	0.21
Wheat	**0.00**	**0.00**	**0.00**	0.19
Temperature	0.54	0.83	0.68	0.75
Internet	**0.00**	**0.01**	0.05	**0.01**
Call centre	**0.02**	**0.00**	**0.00**	0.07
Radio	**0.00**	**0.00**	**0.00**	**0.00**
Gas production	**0.00**	**0.01**	**0.01**	0.11
Sulphuric	0.29	0.38	0.34	0.52
Unemployment	**0.00**	**0.02**	**0.00**	**0.01**
Births	**0.00**	**0.02**	**0.00**	0.12
Wages	**0.00**	**0.01**	**0.01**	**0.00**

Table 1.1: Two sample test p-values applied to 13 time series and 4 regression algorithms. Bold values indicate a positive discovery using a Benjamini–Hochberg procedure with a false discovery rate of 0.05 for each model construction method.

function.

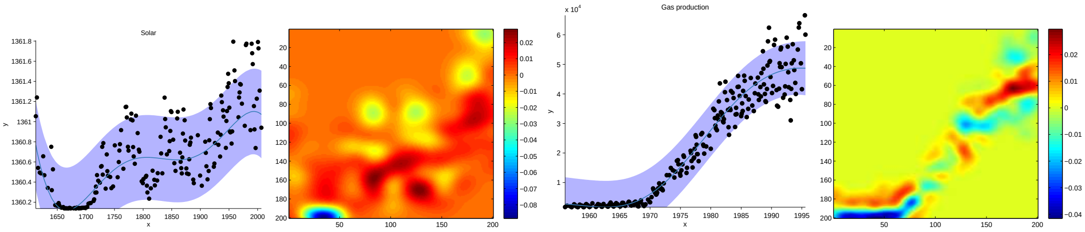

Figure 1.4: From left to right. Solar data with SE posterior. Witness function of SE fit to solar. Gas production data with SE posterior. Witness function of SE fit to gas production.

Of the four methods compared here, only ABCD is able to model heteroscedasticity, explaining why it is the only method with a substantially different set of significant p-values. However, the procedure is still potentially failing to capture structure on four of the datasets.

Figure 1.5 shows the unemployment and Internet data sets, the posterior distribution for the ABCD fits to the data and the witness functions of the ABCD fits. The ABCD method has captured much of the structure in these data sets, making it difficult to visually identify discrepancies between model and data. The witness function for unemployment shows peaks and troughs at similar values of the input x. Comparing to the raw data we see that at these input values there are consistent outliers. Since ABCD is based on Gaussianity assumptions these consistent outliers have caused the method to estimate a large variance in this region, when the true data is non-Gaussian. There is also a similar pattern of peaks and troughs on the internet data suggesting that non-normality has again been detected. Indeed, the data appears to have a strict lower bound which is inconsistent with Gaussianity.

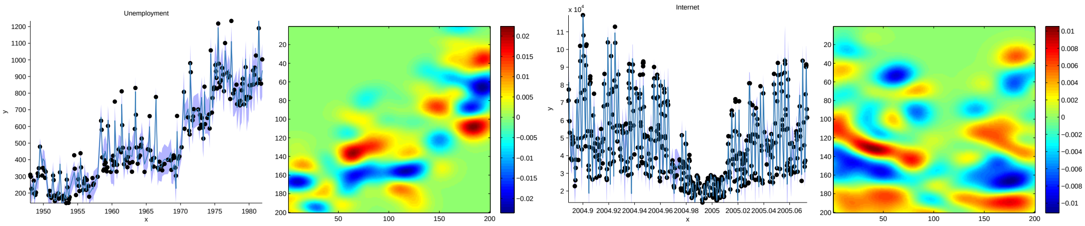

Figure 1.5: From left to right. Unemployment data with ABCD posterior. Witness function of ABCD fit to unemployment. Internet data with ABCD posterior. Witness function of ABCD fit to Internet.

1.7 Discussion of model criticism and related work

1.7.1 Are we criticising a particular model, or class of models?

In section 1.2 we interpreted the differences between classical, Bayesian prior/posterior and plug-in p-values as corresponding to different null hypotheses and interpretations of the word 'model'. In particular the classical p-value tests a null hypothesis that the data could have been generated by a class of distributions (e.g. all normal distributions) whereas all other p-values test a particular probability distribution.

Robins, van der Vaart & Ventura (?) demonstrated that Bayesian and plug-in p-values are not classical p-values (frequentist p-values in their terminology) i.e. they do not have a uniform distribution under the relevant null hypothesis. However, this was presented as a failure of these methods; in particular they demonstrated that methods proposed by Bayarri & Berger (?) based on posterior predictive p-values are asymptotically classical p-values.

This claimed inadequacy of posterior predictive p-values was rebutted (?) and while their usefulness is becoming more accepted (see e.g. introduction of ?) it would appear there is still confusion on the subject (?). We hope that our interpretation of the differences between these methods as different null hypotheses — appropriate in different circustances — sheds further light on the matter.

1.7.2 Should we worry about using the same data for traning and criticism?

Plug-in and posterior predictive p-values test the null hypothesis that the observed data could have been generated by the fitted model or posterior predictive distribution. In some situations it may be more appropriate to attempt to falsify the null hypothesis that future data will be generated by the plug-in or posterior predictive distribution. As mentioned in section 1.2 this can be achieved by reserving a portion of the data to be used for model criticism alone, rather than fitting a model or updating a posterior on the full data. Cross validation methods have also been investigated in this context (???).

1.7.3 Omnibus methods for model criticism

Gelman, Meng & Stern (?) proposed an extension to posterior predictive p-values by generalising the test statistic to a discrepancy measure that can also take model parameters as input as well as data. They also proposed an omnibus test discrepancy measure

$$X^2(Y, \theta) = \sum_{i=1}^{n} \frac{(y_i - \mathbb{E}(y_i \,|\, \theta))^2}{\mathrm{Var}(y_i \,|\, \theta)} \tag{1.9}$$

which resembles a classical χ^2 statistic. While this statistic can be useful in a number of different models, it would have failed to detect any discrepancies in our two toy data examples. In the case of i.i.d. or exchangeable data (and a model that respects this symmetry) this statistic only measures how well the mean and variance of the data has been captured. Since the mean and variance of the data were fit (near) directly by our proposed models we would only expect to see siginificant discrepancies if inference failed.

1.7.4 Other methods for evaluating statistical models

Other typical methods of model evaluation include estimating the predictive performance of the model, analyses of sensitivities to modelling parameters / priors, graphical tests, and estimates of model utility. For a recent survey of Bayesian methods for model assessment, selection and comparison see ? which phrases many techniques as estimates of the utility of a model. For some discussion of sensitivity analysis and graphical model comparison see (e.g. ?).

In this manuscript we have focused on methods that compare statistics of data with predictive distributions, ignoring parameters of the model. When working with models in which individual parameters are of interest (e.g. hierarchical models) other techniques are relevant. O'Hagan (?) proposes a method and selectively reviews techniques appropriate in this context. The discrepancy measures of ? are also relevant.

In the spirit of scientific falsification (e.g. ?), ideally all methods of assessing a model should be performed to gain confidence in any conclusions made. Of course, when performing multiple hypothesis tests care must be taken in the intrepetation of individual p-values.

1.8 Conclusions and future work

In this paper we have demonstrated an exploratory form of model criticism based on two sample tests using kernel maximum mean discrepancy. In contrast to other methods for model criticism, the test analytically maximises over a broad class of statistics, automatically identifying the statistic which most demonstrates the discrepancy between the model and data. We demonstrated how this method of model criticism can be applied to neural networks and Gaussian process regression and demonstrated the ways in which these models were misrepresenting the data they were trained on.

We have demonstrated how kernel MMD two sample tests can be applied to model criticism, but they can be applied to any aspect of statistical modelling where two sample tests are appropriate. This includes for example, Geweke's tests of markov chain posterior sampler validity (?) and tests of markov chain convergence (e.g. ?).

The two sample tests proposed in this paper naturally apply to i.i.d. data and models, but model criticism techniques should of course apply to models with other symmetries (e.g. exchangeable data, logitudinal data / time series, graphs, functions any many others). We have demonstrated an adaptation of the kernel MMD test to regression models However, it is unclear whether maximum mean discrepancy measures can be naturally extended to all of these classes; investigating such extensions would be a profitable area for future study.

In proposing a new method of model criticism we hope we have also exposed the machine learning community to a useful set of tools for diagnosing potential inadequacies of models. On this note, we conclude with a question. Do you know how the model you are currently working with most misrepresents the data it is attempting to model?