Statistical model criticism using kernel two sample tests

James Robert Lloyd

Department of Engineering, University of Cambridge, UK

October 4, 2014

WHY CHECK OR CRITICISE MODELS?

- Statistical analyses are based on assumptions…
  - e.g. Linearity, Gaussianity, Stationarity etc.

WHY CHECK OR CRITICISE MODELS?

- Statistical analyses are based on assumptions…
  - e.g. Linearity, Gaussianity, Stationarity etc.

- …but reality typically breaks these assumptions…
  - ‘A man in daily muddy contact with field experiments could not be expected to have much faith in any direct assumption of independently distributed normal errors’ [Box76]

WHY CHECK OR CRITICISE MODELS?

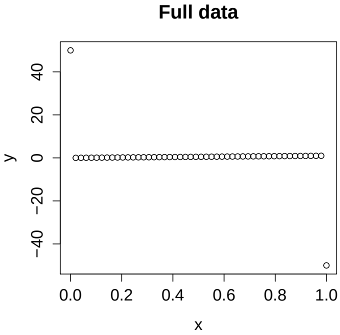
- Statistical analyses are based on assumptions…
  - e.g. Linearity, Gaussianity, Stationarity etc.
- …but reality typically breaks these assumptions…
  - 'A man in daily muddy contact with field experiments could not be expected to have much faith in any direct assumption of independently distributed normal errors' [Box76]
- …and this can lead us to produce false inferences
  - 'We were seeing things that were 25-standard deviation moves, several days in a row'

EXAMPLE: LINEAR REGRESSION

```
Call:
lm(formula = y ~ x)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    5.802      2.702   2.148   0.0368 *
x            -10.645      4.656  -2.286   0.0267 *
```

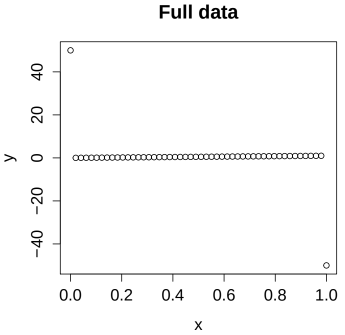
EXAMPLE: LINEAR REGRESSION

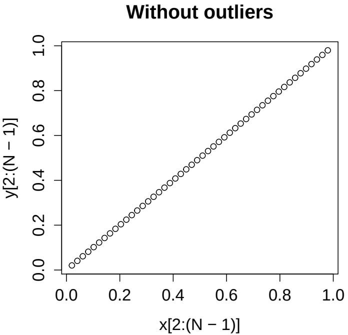
```
Call:
lm(formula = y ~ x)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.802      2.702   2.148   0.0368 *
x           -10.645      4.656  -2.286   0.0267 *
```



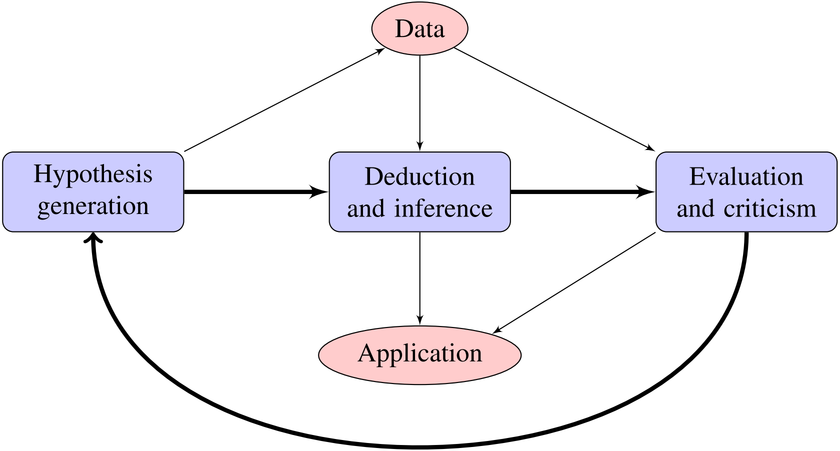
Example: linear regression

```
Call:
lm(formula = y ~ x)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    5.802      2.702   2.148   0.0368 *
x            -10.645      4.656  -2.286   0.0267 *
```

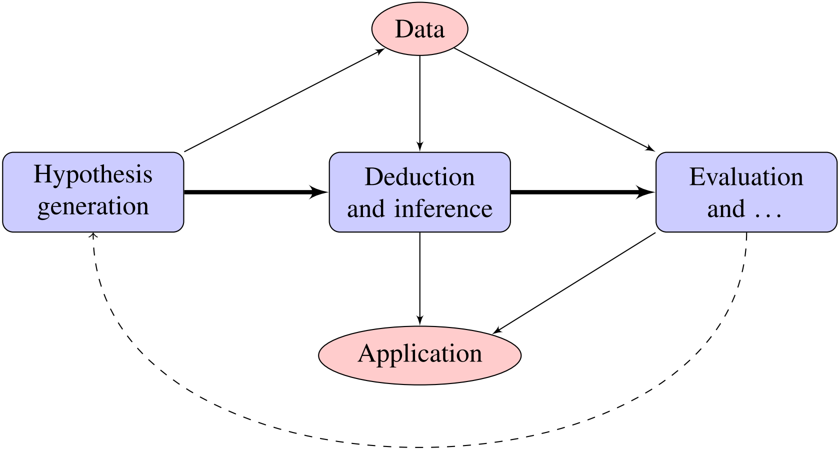




A version of the scientific method



MANY ML PAPERS STOP AT EVALUATION



AGENDA

- Why I became interested in model criticism
- Review of frequentist and Bayesian theory
- A concern about calibration and a potential resolution
- An application of a nonparametric test to model criticism
- Discussion

MY JOURNEY WITH MODEL CRITICISM

- My previous research has involved automatic statistical model building

MY JOURNEY WITH MODEL CRITICISM

- My previous research has involved automatic statistical model building

- I wanted these model building systems to know when they had produced a model which was ‘obviously’ wrong

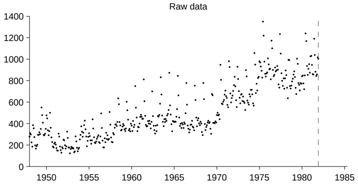
My journey with model criticism

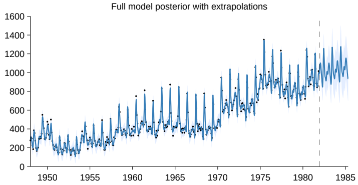
- My previous research has involved automatic statistical model building
- I wanted these model building systems to know when they had produced a model which was 'obviously' wrong
- On entering the literature I found a Bayesians vs frequentists debate that does not appear to have been resolved…

My journey with model criticism

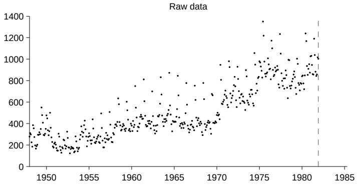
- My previous research has involved automatic statistical model building
- I wanted these model building systems to know when they had produced a model which was 'obviously' wrong
- On entering the literature I found a Bayesians vs frequentists debate that does not appear to have been resolved…
- …and generally very little actionable advice on which method of model criticism to use and when

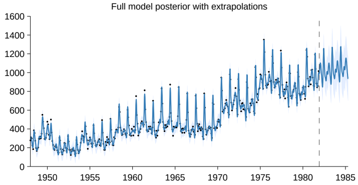
IS THIS MODEL ‘CORRECT’?

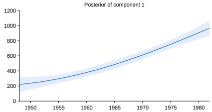


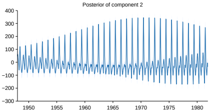


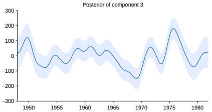
IS THIS MODEL 'CORRECT'?

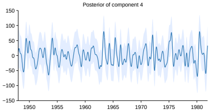


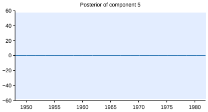












Frequentist model criticism

- We suppose that our data are generated by some parametric model with unknown parameters
  - $X \mid \theta \sim f(x \mid \theta)$

Frequentist model criticism

- We suppose that our data are generated by some parametric model with unknown parameters
  - $X \mid \theta \sim f(x \mid \theta)$

- We wish to test this null hypothesis and control the rate of false positives (Type I errors)

Frequentist model criticism

- We suppose that our data are generated by some parametric model with unknown parameters
  - $X \mid \theta \sim f(x \mid \theta)$
- We wish to test this null hypothesis and control the rate of false positives (Type I errors)
- A typical method is to calculate a *p*-value which is a function of the data

Frequentist model criticism

- We suppose that our data are generated by some parametric model with unknown parameters
  - $X \mid \theta \sim f(x \mid \theta)$
- We wish to test this null hypothesis and control the rate of false positives (Type I errors)
- A typical method is to calculate a p-value which is a function of the data
- A frequentist p-value is a random variable which has a uniform $[0, 1]$ distribution under the null hypothesis

FREQUENTIST MODEL CRITICISM

- Constructing a p-value typically proceeds by choosing a statistic T which is a function of the data
  - The statistic is chosen such that large values are undesirable e.g. outliers

- Constructing a p-value typically proceeds by choosing a statistic T which is a function of the data
  - The statistic is chosen such that large values are undesirable e.g. outliers

- If θ were known then we could define p-values as
  - $p(x_{\text{obs}}) = \mathbb{P}_{f(x \mid \theta)}(T(X) > T(x_{\text{obs}}))$
  - i.e. We will be suspicious of our model when $T(x_{\text{obs}})$ is large

Frequentist model criticism

- Constructing a p-value typically proceeds by choosing a statistic T which is a function of the data
  - The statistic is chosen such that large values are undesirable e.g. outliers

- If θ were known then we could define p-values as
  - $p(x_{\text{obs}}) = \mathbb{P}_{f(x \mid \theta)}(T(X) > T(x_{\text{obs}}))$
  - i.e. We will be suspicious of our model when $T(x_{\text{obs}})$ is large

- When θ is unknown a typical approach is to choose T such that its distribution is independent of θ
  - These are called pivotal quantites
  - Studentised residuals are an example of a pivotal quantity

Maximum likelihood linear regression

- Assume that outputs y are linearly related to inputs X plus independent Gaussian errors or noise ε
  - $y = X\beta + \varepsilon$
  - $\varepsilon \sim_{\text{iid}} \mathcal{N}(0, \sigma^2)$

Maximum likelihood linear regression

- Assume that outputs y are linearly related to inputs X plus independent Gaussian errors or noise ε
  - $y = X\beta + \varepsilon$
  - $\varepsilon \sim_{\text{iid}} \mathcal{N}(0, \sigma^2)$

- Maximum likelihood solution can be found analytically
  - $\hat{\beta} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y$
  - $\hat{\sigma}^2 = \|y - X\hat{\beta}\|^2/n$

Maximum likelihood linear regression

- Assume that outputs y are linearly related to inputs X plus independent Gaussian errors or noise ε
  - $y = X\beta + \varepsilon$
  - $\varepsilon \sim_{\text{iid}} \mathcal{N}(0, \sigma^2)$

- Maximum likelihood solution can be found analytically
  - $\hat{\beta} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y$
  - $\hat{\sigma}^2 = \|y - X\hat{\beta}\|^2/n$

- Many assumptions to be tested before we believe these solutions
  - e.g. X non-random, linearity, constant variance, Gaussianity, independent errors

OUTLIERS

- Perhaps some of the errors are unexpectedly large?

OUTLIERS

- Perhaps some of the errors are unexpectedly large?
- We can check this by comparing $\hat{\varepsilon} = y - X\hat{\beta}$ to its expected distribution

OUTLIERS

- Perhaps some of the errors are unexpectedly large?
- We can check this by comparing $\hat{\varepsilon} = y - X\hat{\beta}$ to its expected distribution
- Let $H = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$, then $\frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$ has a standard t-distribution when $\hat{\sigma}$ is the standard unbiased estimator of σ

OUTLIERS

- Perhaps some of the errors are unexpectedly large?
- We can check this by comparing $\hat{\varepsilon} = y - X\hat{\beta}$ to its expected distribution
- Let $H = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$, then $\frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$ has a standard t-distribution when $\hat{\sigma}$ is the standard unbiased estimator of σ
- These are the studentised residuals
  - They are pivotal quantities since their distribution does not depend on β or σ.

BAYESIAN MODEL CRITICISM

- The frequentist approach tests whether or not the data x_{obs} could have been generated by the 'model' $f(x \mid \theta)$ for any θ

- The frequentist approach tests whether or not the data x_{obs} could have been generated by the 'model' $f(x \mid \theta)$ for any θ
- A more relevant null hypothesis for a Bayesian is that the data was generated from the prior predictive distribution
  - Prior predictive just means the prior over data (as opposed to parameters)

Bayesian model criticism

- The frequentist approach tests whether or not the data x_{obs} could have been generated by the 'model' $f(x \mid \theta)$ for any θ

- A more relevant null hypothesis for a Bayesian is that the data was generated from the prior predictive distribution
  - Prior predictive just means the prior over data (as opposed to parameters)

- We could therefore compute prior predictive p-values
  - $p_{\text{prior}}(x_{\text{obs}}) = \mathbb{P}_{f(x \mid \theta)\pi(\theta)}(T(X) > T(x_{\text{obs}}))$
  - [Box80]

Prior predictive p-values

- These p-values allow us to answer the question:
  - Is some aspect of the data extreme given my prior assumptions?

PRIOR PREDICTIVE p-VALUES

- These p-values allow us to answer the question:
  - Is some aspect of the data extreme given my prior assumptions?
- The statistic T measures the way in which the data is extreme

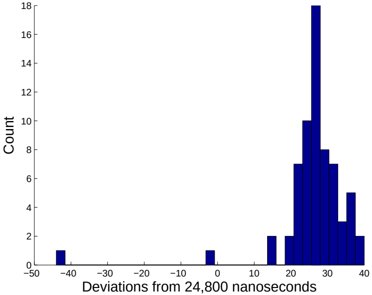
Prior predictive p-values

- These p-values allow us to answer the question:
  - Is some aspect of the data extreme given my prior assumptions?
- The statistic T measures the way in which the data is extreme
- Ill-defined when using improper priors

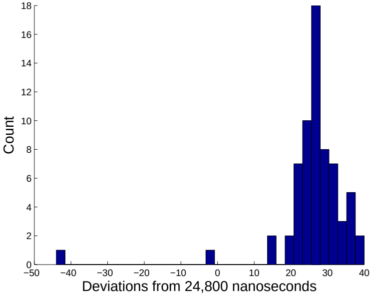
PRIOR PREDICTIVE p-VALUES

- These p-values allow us to answer the question:
  - Is some aspect of the data extreme given my prior assumptions?
- The statistic T measures the way in which the data is extreme
- Ill-defined when using improper priors
- Vague priors can lead to vague tests
  - Probably best used when one really has a subjective prior or the model is not vague about the test statistic

EXAMPLE: SPEED OF LIGHT DATA

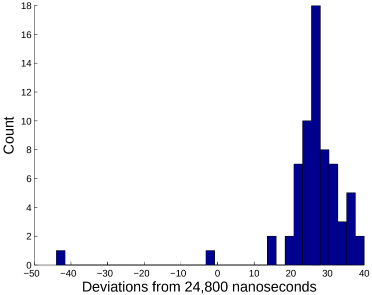


Example: Speed of light data



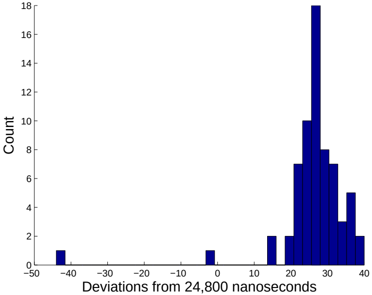
- Suppose we wish to fit a normal distribution to this data with vague priors on the mean and variance

Example: Speed of light data



- Suppose we wish to fit a normal distribution to this data with vague priors on the mean and variance
- We could test this assumption using skewness as the test statistic

EXAMPLE: SPEED OF LIGHT DATA



- Suppose we wish to fit a normal distribution to this data with vague priors on the mean and variance
- We could test this assumption using skewness as the test statistic
  - Skewness = -4.5; *p*-value = tiny

- Rubin [Rub84] instead proposed comparing statistics to their distribution under the posterior distribution
  - $p_{\text{post}}(x_{\text{obs}}) = \mathbb{P}_{f(x \mid \theta)\pi(\theta \mid x_{\text{obs}})}(T(X) > T(x_{\text{obs}}))$

POSTERIOR PREDICTIVE p-VALUES

- Rubin [Rub84] instead proposed comparing statistics to their distribution under the posterior distribution
  - $p_{\text{post}}(x_{\text{obs}}) = \mathbb{P}_{f(x \mid \theta)\pi(\theta \mid x_{\text{obs}})}(T(X) > T(x_{\text{obs}}))$
- These p-values allow us to answer the question:
  - If I were to observe more data, would I be surprised if it was as extreme as the data I originally observed?

POSTERIOR PREDICTIVE p-VALUES

- Rubin [Rub84] instead proposed comparing statistics to their distribution under the posterior distribution
  - $p_{\text{post}}(x_{\text{obs}}) = \mathbb{P}_{f(x \mid \theta)\pi(\theta \mid x_{\text{obs}})}(T(X) > T(x_{\text{obs}}))$
- These p-values allow us to answer the question:
  - If I were to observe more data, would I be surprised if it was as extreme as the data I originally observed?
- One may be concerned about using the data twice
  - The common retort is that the posterior predictive p-value is a well defined subjective probability statement and should be interpreted as such

EXAMPLE: SPEED OF LIGHT DATA AGAIN

- Is the minimum value in the data surprising?

EXAMPLE: SPEED OF LIGHT DATA AGAIN

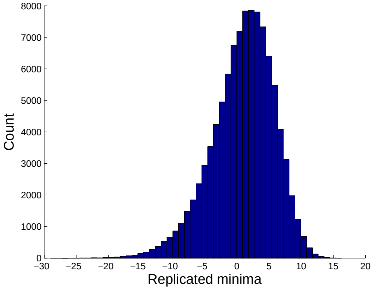
- Is the minimum value in the data surprising?
- A prior predictive would be vague about this quantity

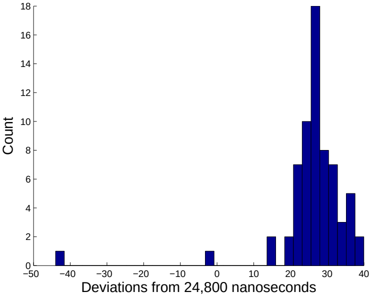
EXAMPLE: SPEED OF LIGHT DATA AGAIN

- Is the minimum value in the data surprising?
- A prior predictive would be vague about this quantity
- Instead we can test this by sampling data sets of the same size from the posterior, and recording their minima

Example: Speed of light data again

- Is the minimum value in the data surprising?
- A prior predictive would be vague about this quantity
- Instead we can test this by sampling data sets of the same size from the posterior, and recording their minima





A PROBLEM OF CALIBRATION

- The Bayesian p-values defined thus far are not frequentist p-values

A PROBLEM OF CALIBRATION

- The Bayesian p-values defined thus far are not frequentist p-values
- Alternatives were proposed by [BB99] and were shown to be asymptotically frequentist p-values by [RvdVV00]

A problem of calibration

- The Bayesian p-values defined thus far are not frequentist p-values
- Alternatives were proposed by [BB99] and were shown to be asymptotically frequentist p-values by [RvdVV00]
- But maybe this isn't a problem — perhaps we are confusing the use of the word 'model'
  - *"If our goal is to check the model $f(x;\theta)$ rather than the prior $\pi(\theta)$, our procedures should perform adequately whatever the prior, including point-mass priors"* — James Robins discussion of [BB99]

DIFFERENT TESTS FOR DIFFERENT GOALS

- Frequentist p-values test an entire class of models
  - The null hypothesis is that the data could have been generated by $f(x \mid \theta)$ for some θ

Different tests for different goals

- Frequentist p-values test an entire class of models
  - The null hypothesis is that the data could have been generated by $f(x \mid \theta)$ for some θ

- Prior predictive p-values test a particular prior distribution
  - The null hypothesis is that the data could have been generated from the prior distribution

DIFFERENT TESTS FOR DIFFERENT GOALS

- Frequentist p-values test an entire class of models
  - The null hypothesis is that the data could have been generated by $f(x \mid \theta)$ for some θ
- Prior predictive p-values test a particular prior distribution
  - The null hypothesis is that the data could have been generated from the prior distribution
- Posterior predictive p-values use the data twice so require a different interpretation
  - Can be interpreted literally as the probability of a future data set being more extreme than the one just observed

DIFFERENT TESTS FOR DIFFERENT GOALS

- Frequentist p-values test an entire class of models
  - The null hypothesis is that the data could have been generated by $f(x \mid \theta)$ for some θ
- Prior predictive p-values test a particular prior distribution
  - The null hypothesis is that the data could have been generated from the prior distribution
- Posterior predictive p-values use the data twice so require a different interpretation
  - Can be interpreted literally as the probability of a future data set being more extreme than the one just observed
- Can also use held out data to test posterior distributions

How do we choose the test statistic?

- All of the tests so far have required a statistic by which one measures if the data is extreme

How do we choose the test statistic?

- All of the tests so far have required a statistic by which one measures if the data is extreme
- We could apply a battery of statistics to every problem
  - Without understanding dependencies between all statistics multiple comparisons adjustments will become highly conservative

HOW DO WE CHOOSE THE TEST STATISTIC?

- All of the tests so far have required a statistic by which one measures if the data is extreme
- We could apply a battery of statistics to every problem
  - Without understanding dependencies between all statistics multiple comparisons adjustments will become highly conservative
- Can we instead define a statistic in high level terms, and then compute the statistic which most demonstrates any discrepancy?

How do we choose the test statistic?

- All of the tests so far have required a statistic by which one measures if the data is extreme
- We could apply a battery of statistics to every problem
  - Without understanding dependencies between all statistics multiple comparisons adjustments will become highly conservative
- Can we instead define a statistic in high level terms, and then compute the statistic which most demonstrates any discrepancy?
  - And will we be able to interpret this statistic?

Maximum mean discrepancy two sample tests

- Suppose we have samples $x \sim_{\text{iid}} p$ and $y \sim_{\text{iid}} q$ and we wish to test the hypothesis $p = q$

Maximum mean discrepancy two sample tests

- Suppose we have samples $x \sim_{\text{iid}} p$ and $y \sim_{\text{iid}} q$ and we wish to test the hypothesis $p = q$

- Define $\text{MMD}(\mathcal{F}, p, q) = \sup_{f \in \mathcal{F}}(\mathbb{E}_{x \sim p}[f(x)] - \mathbb{E}_{y \sim q}[f(y)])$ where $\mathcal{F}$ is a reproducing kernel Hilbert space (RKHS) with kernel k

Maximum mean discrepancy two sample tests

- Suppose we have samples $x \sim_{\text{iid}} p$ and $y \sim_{\text{iid}} q$ and we wish to test the hypothesis $p = q$
- Define $\text{MMD}(\mathcal{F}, p, q) = \sup_{f \in \mathcal{F}} (\mathbb{E}_{x \sim p}[f(x)] - \mathbb{E}_{y \sim q}[f(y)])$ where $\mathcal{F}$ is a reproducing kernel Hilbert space (RKHS) with kernel k
- The function attaining this supremum can be computed analytically
  - $f(x) = \mathbb{E}_{x' \sim p}[k(x, x')] - \mathbb{E}_{x' \sim q}[k(x, x')]$

Maximum mean discrepancy two sample tests

- Suppose we have samples $x \sim_{\text{iid}} p$ and $y \sim_{\text{iid}} q$ and we wish to test the hypothesis $p = q$

- Define $\text{MMD}(\mathcal{F}, p, q) = \sup_{f \in \mathcal{F}}(\mathbb{E}_{x \sim p}[f(x)] - \mathbb{E}_{y \sim q}[f(y)])$ where $\mathcal{F}$ is a reproducing kernel Hilbert space (RKHS) with kernel k

- The function attaining this supremum can be computed analytically
  - $f(x) = \mathbb{E}_{x' \sim p}[k(x, x')] - \mathbb{E}_{x' \sim q}[k(x, x')]$

- Substituting and squaring:
  - $\text{MMD}^2(\mathcal{F}, p, q) = \mathbb{E}_{x,x' \sim p}[k(x, x')] + 2\mathbb{E}_{x \sim p, y \sim q}[k(x, y)] + \mathbb{E}_{y,y' \sim q}[k(y, y')]$

EMPIRICAL ESTIMATION

- We can estimate these expectations from finite samples
  - $\mathrm{MMD}_b^2(\mathcal{F}, X, Y) = \frac{1}{m^2}\sum_{i,j=1}^{m} k(x_i, x_j) - \frac{2}{mn}\sum_{i,j=1}^{m,n} k(x_i, y_j) + \frac{1}{n^2}\sum_{i,j=1}^{n} k(y_i, y_j)$
  - $\hat{f}(x) = \frac{1}{m}\sum_{i=1}^{m} k(x, x_i) - \frac{1}{n}\sum_{i=1}^{n} k(x, y_i)$

EMPIRICAL ESTIMATION

- We can estimate these expectations from finite samples
  - $\text{MMD}_b^2(\mathcal{F}, X, Y) =$
    $\frac{1}{m^2}\sum_{i,j=1}^{m} k(x_i, x_j) - \frac{2}{mn}\sum_{i,j=1}^{m,n} k(x_i, y_j) + \frac{1}{n^2}\sum_{i,j=1}^{n} k(y_i, y_j)$
  - $\hat{f}(x) = \frac{1}{m}\sum_{i=1}^{m} k(x, x_i) - \frac{1}{n}\sum_{i=1}^{n} k(x, y_i)$

- The empirical witness function is just the difference of two kernel density estimates

EMPIRICAL ESTIMATION

- We can estimate these expectations from finite samples
  - $\mathrm{MMD}_b^2(\mathcal{F}, X, Y) =$
    $\frac{1}{m^2}\sum_{i,j=1}^{m} k(x_i, x_j) - \frac{2}{mn}\sum_{i,j=1}^{m,n} k(x_i, y_j) + \frac{1}{n^2}\sum_{i,j=1}^{n} k(y_i, y_j)$
  - $\hat{f}(x) = \frac{1}{m}\sum_{i=1}^{m} k(x, x_i) - \frac{1}{n}\sum_{i=1}^{n} k(x, y_i)$
- The empirical witness function is just the difference of two kernel density estimates
- We can estimate the null distribution of the MMD statistic by a bootstrap procedure
  - It is an example of a permutation test

APPLICATION TO MODEL CHECKING

- Need to make some simplifying assumptions
  - Data y are generated $i.i.d.$ from some distribution q
  - We make a point estimate of q which we denote p

APPLICATION TO MODEL CHECKING

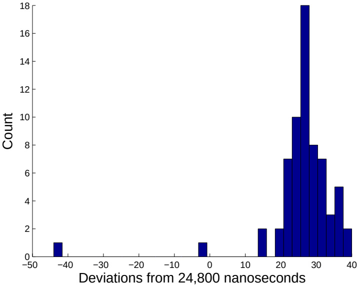
- Need to make some simplifying assumptions
  - Data y are generated $i.i.d.$ from some distribution q
  - We make a point estimate of q which we denote p

- The hypothesis that $p = q$ is now the hypothesis that the model is correct

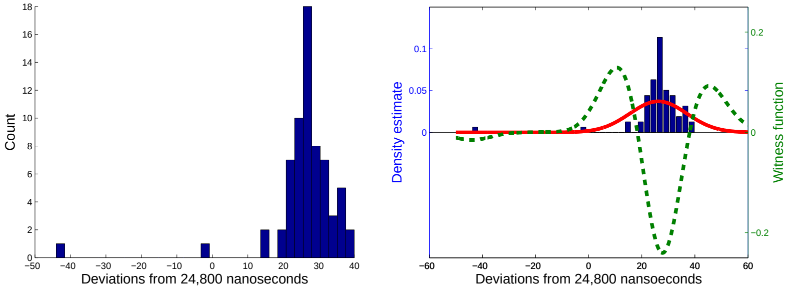
Application to model checking

- Need to make some simplifying assumptions
  - Data y are generated $i.i.d.$ from some distribution q
  - We make a point estimate of q which we denote p
- The hypothesis that $p = q$ is now the hypothesis that the model is correct
- We can generate samples from p and then perform a two sample test

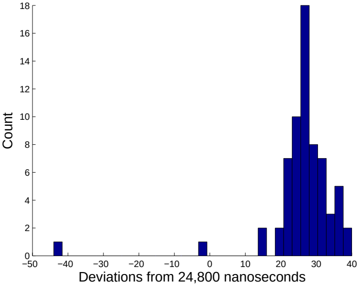
EXAMPLE: SPEED OF LIGHT DATA AGAIN (AGAIN)

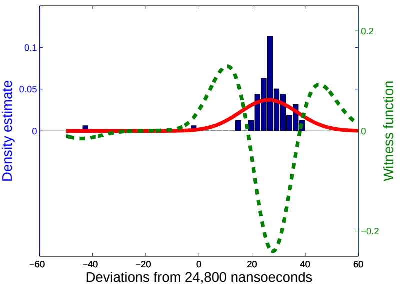


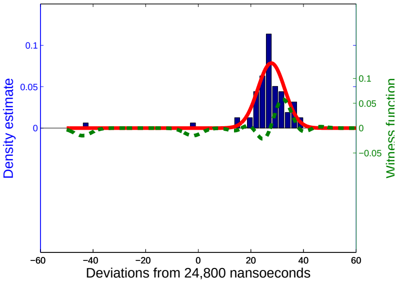
EXAMPLE: SPEED OF LIGHT DATA AGAIN (AGAIN)



EXAMPLE: SPEED OF LIGHT DATA AGAIN (AGAIN)







What happens in high dimensions?

- Interpretability of test comes from interpretation of witness function as difference of kernel density estimates

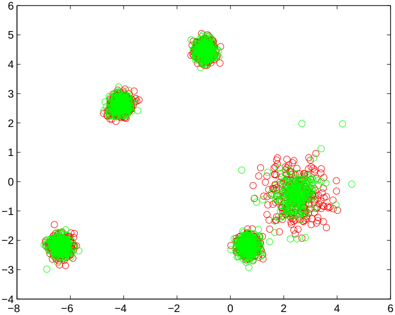
What happens in high dimensions?

- Interpretability of test comes from interpretation of witness function as difference of kernel density estimates
- Kernel density estimation is high variance in high dimensions and will likely be uninterpretable

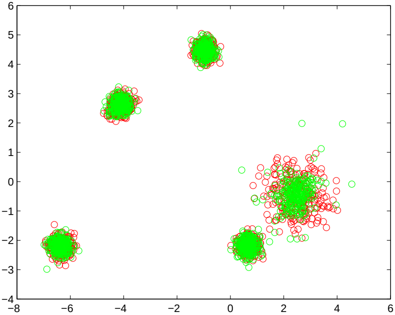
What happens in high dimensions?

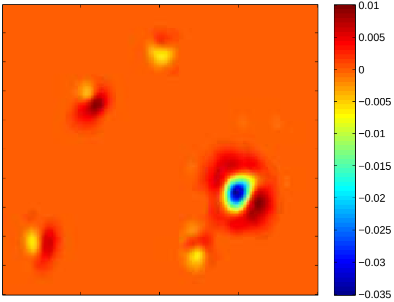
- Interpretability of test comes from interpretation of witness function as difference of kernel density estimates
- Kernel density estimation is high variance in high dimensions and will likely be uninterpretable
- Potential solution: Include dimensionality reduction as part of the test statistic
  - $\frac{1}{m^2}\sum_{i,j=1}^{m} k(x_i^{\text{PCA}}, x_j^{\text{PCA}}) - \frac{2}{mn}\sum_{i,j=1}^{m,n} k(x_i^{\text{PCA}}, y_j^{\text{PCA}}) + \frac{1}{n^2}\sum_{i,j=1}^{n} k(y_i^{\text{PCA}}, y_j^{\text{PCA}})$

WHAT HAPPENS IN HIGH DIMENSIONS?

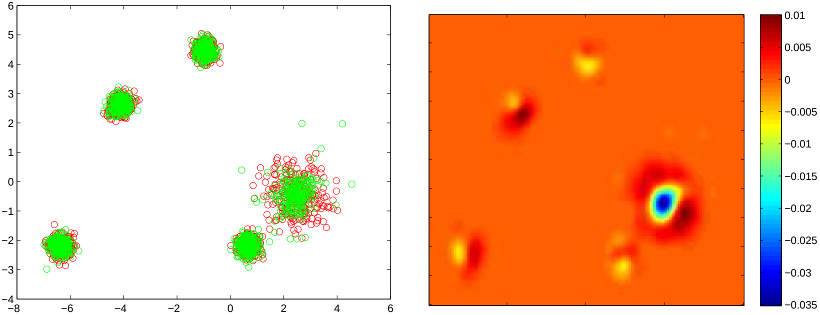


WHAT HAPPENS IN HIGH DIMENSIONS?





What happens in high dimensions?



- Estimated p-value of 0.05

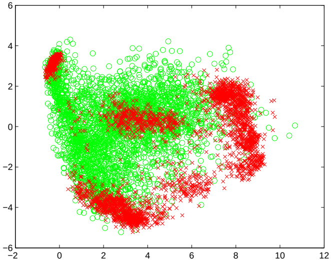
WHAT DO NEURAL NETWORKS DREAM ABOUT?

- Various versions of deep belief networks have been trained to produce generative models of MNIST handwritten digits

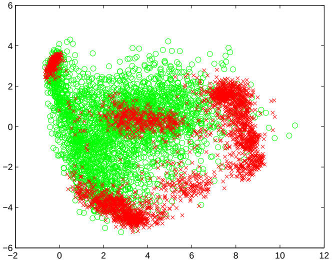
What do neural networks dream about?

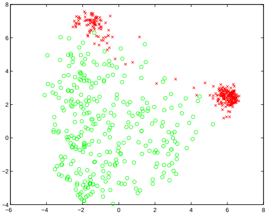
- Various versions of deep belief networks have been trained to produce generative models of MNIST handwritten digits
- Samples from these models certainly look like digits, but what aspects of the distribution over handwritten digits do these models not capture?

AN RBM TRAINED ON MNIST

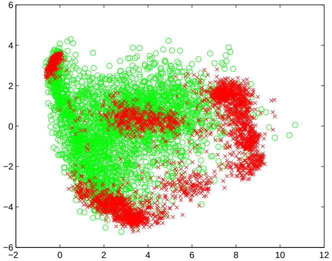


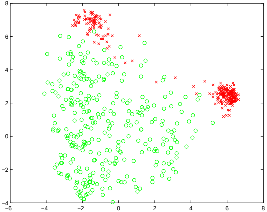
AN RBM TRAINED ON MNIST

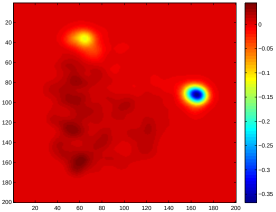




An RBM trained on MNIST

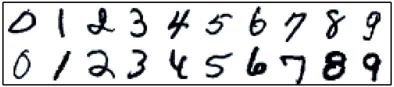




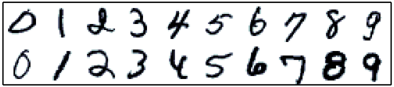


What do neural networks dream about?

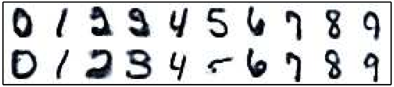
Real digits



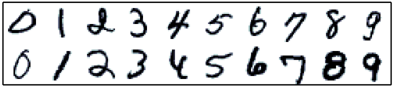
Real digits



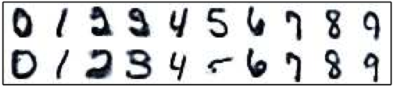
Witness fn troughs : RBM



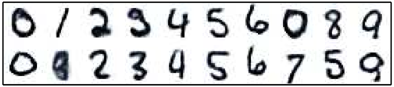
Real digits



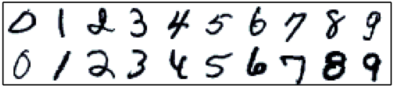
Witness fn troughs : RBM



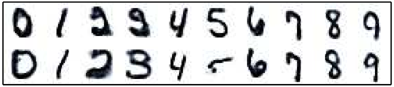
Witness fn troughs : RBMs



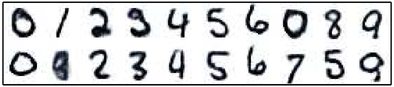
Real digits



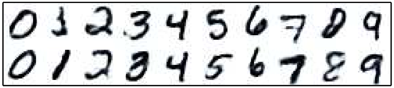
Witness fn troughs : RBM



Witness fn troughs : RBMs



Witness fn troughs : DBN



WHY HAVE I NOT MENTIONED POWER?

- Quantification of power requires specification of an explicit alternative model or hypothesis

WHY HAVE I NOT MENTIONED POWER?

- Quantification of power requires specification of an explicit alternative model or hypothesis
- "Model criticism . . . is intended as an open-minded phase of investigation to identify any problems with the model. Formulation of explicit alternatives comes after the model criticism phase has identified some problems." [O'H03]

Why have I not mentioned power?

- Quantification of power requires specification of an explicit alternative model or hypothesis
- "Model criticism ... is intended as an open-minded phase of investigation to identify any problems with the model. Formulation of explicit alternatives comes after the model criticism phase has identified some problems." [O'H03]
- I'm not sure anymore that alternative free hypothesis tests are the correct way to think about model criticism
  - Some effort should be made to characterise the alternative hypotheses for which certain tests have high power

FUTURE WORK

- Can we usefully characterise the types of alternative models for which MMD tests have high power?

- Can we usefully characterise the types of alternative models for which MMD tests have high power?
  - Using different kernels will identify different types of discrepancy

FUTURE WORK

- Can we usefully characterise the types of alternative models for which MMD tests have high power?
  - Using different kernels will identify different types of discrepancy
  - This characterisation probably already in the literature

FUTURE WORK

- Can we usefully characterise the types of alternative models for which MMD tests have high power?
  - Using different kernels will identify different types of discrepancy
  - This characterisation probably already in the literature
- Could these alternatives be usefully understood as specific nonparametric models by exploiting the connections between RKHSs and Gaussian processes?

DISCUSSION

- Box advocated for model criticism to be part of the loop of the scientific process…
  - Probabilistic version of falsification

Discussion

- Box advocated for model criticism to be part of the loop of the scientific process…
  - Probabilistic version of falsification

- …but should we just be estimating the utility of a model?

DISCUSSION

- Box advocated for model criticism to be part of the loop of the scientific process…
  - Probabilistic version of falsification
- …but should we just be estimating the utility of a model?
- Perhaps we should be using model criticism to estimate the utility of expanding a model?
  - They are often much cheaper to compute (thinking and computing time) than performing inference in an expanded model class

Discussion

- Box advocated for model criticism to be part of the loop of the scientific process…
  - Probabilistic version of falsification
- …but should we just be estimating the utility of a model?
- Perhaps we should be using model criticism to estimate the utility of expanding a model?
  - They are often much cheaper to compute (thinking and computing time) than performing inference in an expanded model class
- What are other forms of model criticism that are widely applicable and can help identify the nature of discrepancies between model and data in highly complicated systems

References I

[BB99] MJ Bayarri and JO Berger. Quantifying surprise in the data and model verification. *Bayesian statistics*, 1999.

[Box76] George E P Box. Science and statistics. *J. Am. Stat. Assoc.*, 71(356):791–799, 1 December 1976.

[Box80] George E. P. Box. Sampling and bayes' inference in scientific modelling and robustness. *J. R. Stat. Soc. Ser. A*, 143(4):383–430, 1 January 1980.

[GMS96] Andrew Gelman, Xiao-Li Meng, and Hal Stern. Posterior predictive assessment of model fitness via realized discrepancies. *Stat. Sin.*, 6:733–807, 1996.

[O'H03] A O'Hagan. HSSS model criticism. *Highly Structured Stochastic Systems*, pages 423–444, 2003.

[Rub84] Donald B. Rubin. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, 12(4):1151–1172, December 1984.

[RvdVV00] James M. Robins, Aad van der Vaart, and Valerie Venture. Asymptotic distribution of p values in composite null models. *Journal of the American Statistical Association*, 95(452):1143–1156, 2000.

Appendix

Is cross validation enough?

- Machine learning is typically concerned with predictive accuracy…

Is cross validation enough?

- Machine learning is typically concerned with predictive accuracy…
- …which can be estimated by cross validation

Algorithm	CV error (standard error)
Linear regression	4.8 (1.3)
Theil–Sen estimator	2.0 (1.4)

Is cross validation enough?

- Machine learning is typically concerned with predictive accuracy…
- …which can be estimated by cross validation

Algorithm	CV error (standard error)
Linear regression	4.8 (1.3)
Theil–Sen estimator	2.0 (1.4)

- So we're fine, right?

IS CROSS VALIDATION ENOUGH?

- Let's try some new data

- Let's try some new data

Algorithm	CV error (standard error)
Linear regression	2.0 (0.1)
Theil–Sen estimator	1.3 (0.2)

IS CROSS VALIDATION ENOUGH?

- Let's try some new data

Algorithm	CV error (standard error)
Linear regression	2.0 (0.1)
Theil–Sen estimator	1.3 (0.2)

- So outliers again?

IS CROSS VALIDATION ENOUGH?

- Let's try some new data

Algorithm	CV error (standard error)
Linear regression	2.0 (0.1)
Theil–Sen estimator	1.3 (0.2)
5 nearest neighbours	0.1 (0.01)

IS CROSS VALIDATION ENOUGH?

- Let's try some new data

Algorithm	CV error (standard error)
Linear regression	2.0 (0.1)
Theil–Sen estimator	1.3 (0.2)
5 nearest neighbours	0.1 (0.01)

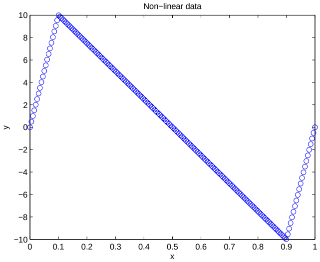
- When I said outliers I actually meant non-linearity

IS CROSS VALIDATION ENOUGH?

- Let's try some new data

Algorithm	CV error (standard error)
Linear regression	2.0 (0.1)
Theil–Sen estimator	1.3 (0.2)
5 nearest neighbours	0.1 (0.01)

- When I said outliers I actually meant non-linearity



Is being Bayesian enough?

- Why not just do all inference in a super model that contains everything we could ever possibly believe?

IS BEING BAYESIAN ENOUGH?

- Why not just do all inference in a super model that contains everything we could ever possibly believe?
  - And then e.g. estimate the utility of smaller models for certain tasks

Is being Bayesian enough?

- Why not just do all inference in a super model that contains everything we could ever possibly believe?
  - And then e.g. estimate the utility of smaller models for certain tasks

- How long did you spend coding your last inference scheme?
  - Or have you ever got probabilistic programming to work in a non-trivial model?

Is being Bayesian enough?

- Why not just do all inference in a super model that contains everything we could ever possibly believe?
  - And then e.g. estimate the utility of smaller models for certain tasks
- How long did you spend coding your last inference scheme?
  - Or have you ever got probabilistic programming to work in a non-trivial model?
- Model criticism / checking gives us tools to explore potential inadequacies of a method…
  - …without having to implement inference for every expanded method we can think of

- Gelman et alia [GMS96] proposed generalising the statistic $T(x)$ of posterior predictive p-values to a discrepancy measure $d(x, \theta)$ which depends on the parameters θ of the model

DISCREPANCY p-VALUES

- Gelman et alia [GMS96] proposed generalising the statistic $T(x)$ of posterior predictive p-values to a discrepancy measure $d(x, \theta)$ which depends on the parameters θ of the model

- The observed discrepancy is again compared to the posterior predictive distribution

- Gelman et alia [GMS96] proposed generalising the statistic $T(x)$ of posterior predictive p-values to a discrepancy measure $d(x, \theta)$ which depends on the parameters θ of the model
- The observed discrepancy is again compared to the posterior predictive distribution
- $p_{\text{dis}}(x_{\text{obs}}) = \mathbb{P}(d(X, \theta) \geq d(x_{\text{obs}}, \theta) \,|\, x_{\text{obs}})$

DISCREPANCY p-VALUES

- Gelman et alia [GMS96] proposed generalising the statistic $T(x)$ of posterior predictive p-values to a discrepancy measure $d(x, \theta)$ which depends on the parameters θ of the model

- The observed discrepancy is again compared to the posterior predictive distribution

- $p_{\text{dis}}(x_{\text{obs}}) = \mathbb{P}(d(X, \theta) \geq d(x_{\text{obs}}, \theta) \,|\, x_{\text{obs}})$

- Can be estimated using samples of the joint posterior distribution of (X, θ) e.g. from MCMC

Example: Forecasting U.S. elections

- Copied without permission from Gelman, A. et al. Bayesian Data Analysis, Third Edition. (Taylor & Francis, 2013)

- Fitting a linear model to proportion of votes for Democrats by state for 11 presidential elections

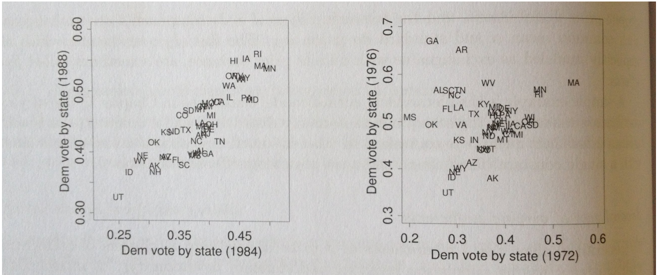


Figure 15.1 *(a) Democratic share of the two-party vote for president, for each state, in 1984 and 1988. (b) Democratic share of the two-party vote for president, for each state, in 1972 and 1976.*

EXAMPLE: FORECASTING U.S. ELECTIONS

Description of variable	Sample quantiles min	median	max
Nationwide variables:			
Support for Dem. candidate in Sept. poll	0.37	0.46	0.69
(Presidential approval in July poll) × Inc	−0.69	−0.47	0.74
(Presidential approval in July poll) × Presinc	−0.69	0	0.74
(2nd quarter GNP growth) × Inc	−0.024	−0.005	0.018
Statewide variables:			
Dem. share of state vote in last election	−0.23	−0.02	0.41
Dem. share of state vote two elections ago	−0.48	−0.02	0.41
Home states of presidential candidates	−1	0	1
Home states of vice-presidential candidates	−1	0	1
Democratic majority in the state legislature	−0.49	0.07	0.50
(State economic growth in past year) × Inc	−0.22	−0.00	0.26
Measure of state ideology	−0.78	−0.02	0.69
Ideological compatibility with candidates	−0.32	−0.05	0.32
Proportion Catholic in 1960 (compared to U.S. avg.)	−0.21	0	0.38
Regional/subregional variables:			
South	0	0	1
(South in 1964) × (−1)	−1	0	0
(Deep South in 1964) × (−1)	−1	0	0
New England in 1964	0	0	1
North Central in 1972	0	0	1
(West in 1976) × (−1)	−1	0	0

Example: Forecasting U.S. elections

- Linear model assumes that each (input, output) tuple is exchangeable

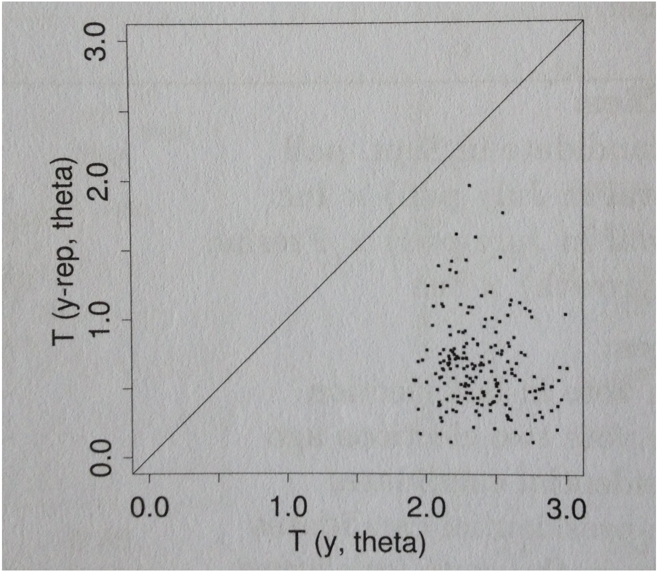
EXAMPLE: FORECASTING U.S. ELECTIONS

- Linear model assumes that each (input, output) tuple is exchangeable
- This ignores any correlation between states in a particular year due to nationwide swings in voting

Example: Forecasting U.S. elections

- Linear model assumes that each (input, output) tuple is exchangeable
- This ignores any correlation between states in a particular year due to nationwide swings in voting
- We can construct a test staistic that can test for this correlation
  - For each year compute the average error of the prediction in each state
  - Compute the root mean square of these averageerrors over the 11 years

EXAMPLE: FORECASTING U.S. ELECTIONS



EXAMPLE: FORECASTING U.S. ELECTIONS

- The posterior distribution underestimates the test statistic

Example: Forecasting U.S. elections

- The posterior distribution underestimates the test statistic
- The practical consequence of this is that the model will give overly precise predictions of national election results

EXAMPLE: FORECASTING U.S. ELECTIONS

- The posterior distribution underestimates the test statistic
- The practical consequence of this is that the model will give overly precise predictions of national election results
- The model can be improved by adding indicator variables for each year
  - Can also include year $\times$ region features to capture regional voting swings

Model criticism for Gaussian processes

- Suppose that $y_i \sim f(x_i) + \varepsilon_i$ where
  - $f \sim \mathcal{GP}(0, k)$
  - $\varepsilon \sim_{\text{iid}} \mathcal{N}(0, \sigma^2)$

- Suppose that $y_i \sim f(x_i) + \varepsilon_i$ where
  - $f \sim \mathcal{GP}(0, k)$
  - $\varepsilon \sim_{\text{iid}} \mathcal{N}(0, \sigma^2)$

- We might be interested in testing the residuals

- Suppose that $y_i \sim f(x_i) + \varepsilon_i$ where
  - $f \sim \mathcal{GP}(0, k)$
  - $\varepsilon \sim_{\text{iid}} \mathcal{N}(0, \sigma^2)$

- We might be interested in testing the residuals
  - We can construct a discrepancy measure based on some function of $y_i - f(x_i)$

Model criticism for Gaussian processes

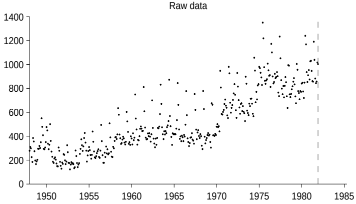
- Suppose that $y_i \sim f(x_i) + \varepsilon_i$ where
  - $f \sim \mathcal{GP}(0, k)$
  - $\varepsilon \sim_{\text{iid}} \mathcal{N}(0, \sigma^2)$
- We might be interested in testing the residuals
  - We can construct a discrepancy measure based on some function of $y_i - f(x_i)$
  - A posterior predictive check would then correspond to comparing some function of the prior and posterior of ε

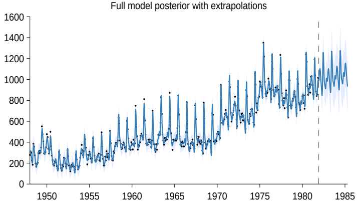
- Suppose that $y_i \sim f(x_i) + \varepsilon_i$ where
  - $f \sim \mathcal{GP}(0, k)$
  - $\varepsilon \sim_{\text{iid}} \mathcal{N}(0, \sigma^2)$
- We might be interested in testing the residuals
  - We can construct a discrepancy measure based on some function of $y_i - f(x_i)$
  - A posterior predictive check would then correspond to comparing some function of the prior and posterior of ε
- Similarly, we could construct discrepancies based on $y_i - \varepsilon_i$

Model criticism for Gaussian processes

- Suppose that $y_i \sim f(x_i) + \varepsilon_i$ where
  - $f \sim \mathcal{GP}(0, k)$
  - $\varepsilon \sim_{\text{iid}} \mathcal{N}(0, \sigma^2)$
- We might be interested in testing the residuals
  - We can construct a discrepancy measure based on some function of $y_i - f(x_i)$
  - A posterior predictive check would then correspond to comparing some function of the prior and posterior of ε
- Similarly, we could construct discrepancies based on $y_i - \varepsilon_i$
  - This amounts to comparing the prior and posterior of f

Example: Is this model 'correct'?





- A very smooth monotonically increasing function.
- An approximately periodic function with a period of 1.0 years.
- A smooth function.
- A smooth function.
- Uncorrelated noise.

DIFFERENT STATISTICS FOR EACH COMPONENT

- p-values of several statistics for each model component
- Mea culpa these p-values are unadjusted for multiple comparisons, but they are also uncalibrated (they are conservative)

	ACF		Periodogram		QQ	
#	min	min loc	max	max loc	max	min
1	0.502	0.582	0.341	0.413	0.341	0.679
2	0.802	0.199	0.558	0.630	0.049	0.785
3	0.251	0.475	0.799	0.447	0.534	0.769
4	0.527	0.503	0.504	0.481	0.430	0.616
5	0.493	0.477	0.503	0.487	0.518	0.381

Example: Identifying outliers

The following discrepancies between the prior and posterior distributions for this component have been detected.

- The qq plot has an unexpectedly large positive deviation from equality ($x = y$). This discrepancy has an estimated p-value of 0.049.