Building an automatic statistician

James Robert Lloyd\textsuperscript{1}, David Duvenaud\textsuperscript{1}, Roger Grosse\textsuperscript{2},

Joshua Tenenbaum\textsuperscript{2}, Zoubin Ghahramani\textsuperscript{1}

\textsuperscript{1}: Department of Engineering, University of Cambridge, UK
\textsuperscript{2}: Massachusetts Institute of Technology, USA

August 4, 2014
We live in an era of abundant data

The McKinsey Global Institute claim

“The United States alone faces a shortage of 140,000 to 190,000 people with analytical expertise and 1.5 million managers and analysts with the skills to understand and make decisions based on the analysis of big data.”

Diverse fields increasingly relying on expert statisticians, machine learning researchers and data scientists e.g.

- Computational sciences (e.g. biology, astronomy, …)
- Online advertising
- Quantitative finance
- …
WHAT WOULD AN AUTOMATIC STATISTICIAN DO?

Language of models

Data  Search  Model  Prediction  Report

Evaluation  Checking

Translation
GOALS OF THE AUTOMATIC STATISTICIAN PROJECT

- Provide a set of tools for understanding data that require minimal expert input

- Uncover challenging research problems in e.g.
  - Automated inference
  - Model construction and comparison
  - Data visualisation and interpretation

- Advance the field of machine learning in general
An open-ended language of models
  Expressive enough to capture real-world phenomena...
  ...and the techniques used by human statisticians

A search procedure
  To efficiently explore the language of models

A principled method of evaluating models
  Trading off complexity and fit to data

A procedure to automatically explain the models
  Making the assumptions of the models explicit...
  ...in a way that is intelligible to non-experts
Four additive components have been identified in the data

- A linearly increasing function.
- An approximately periodic function with a period of 1.0 years and with linearly increasing amplitude.
- A smooth function.
- Uncorrelated noise with linearly increasing standard deviation.
DEFINING A LANGUAGE OF MODELS

Language of models

Data → Search → Model → Prediction → Report

Evaluation → Translation → Checking

James Robert Lloyd and Zoubin Ghahramani
Regression consists of learning a function $f : \mathcal{X} \to \mathcal{Y}$ from inputs to outputs from example input / output pairs.

- Language should include simple parametric forms...
  - e.g. Linear functions, Polynomials, Exponential functions

- ...as well as functions specified by high level properties
  - e.g. Smoothness, Periodicity

- Inference should be tractable for all models in language
WE CAN BUILD REGRESSION MODELS WITH
GAUSSIAN PROCESSES

- GPs are distributions over functions such that any finite subset of function evaluations, \((f(x_1), f(x_2), \ldots, f(x_N))\), have a joint Gaussian distribution.

- A GP is completely specified by
  - Mean function, \(\mu(x) = \mathbb{E}(f(x))\)
  - Covariance / kernel function, \(k(x, x') = \text{Cov}(f(x), f(x'))\)
  - Denoted \(f \sim \text{GP}(\mu, k)\)
It is common practice to use a zero mean function since the mean can be marginalised out

- Suppose, \( f(x) \mid a \sim \text{GP}(a \times \mu(x), k(x, x')) \) where \( a \sim \mathcal{N}(0, 1) \)
- Then equivalently, \( f(x) \sim \text{GP}(0, \mu(x)\mu(x') + k(x, x')) \)

We therefore define a language of GP regression models by specifying a language of kernels
Five base kernels

- Squared exp. (SE)
- Periodic (PER)
- Linear (LIN)
- Constant (C)
- White noise (WN)

Encoding for the following types of functions

- Smooth functions
- Periodic functions
- Linear functions
- Constant functions
- Gaussian noise
Two main operations: addition, multiplication

- **LIN × LIN**
  - Quadratic functions

- **LIN + PER**
  - Periodic plus linear trend

- **SE × PER**
  - Locally periodic

- **SE + PER**
  - Periodic plus smooth trend
Assume \( f_1(x) \sim GP(0, k_1) \) and \( f_2(x) \sim GP(0, k_2) \). Define:

\[
f(x) = (1 - \sigma(x))f_1(x) + \sigma(x)f_2(x)
\]

where \( \sigma \) is a sigmoid function between 0 and 1.

Then \( f \sim GP(0, k) \), where

\[
k(x, x') = (1 - \sigma(x)) k_1(x, x') (1 - \sigma(x')) + \sigma(x) k_2(x, x') \sigma(x')
\]

We define the changepoint operator \( k = CP(k_1, k_2) \).
## An Expressive Language of Models

<table>
<thead>
<tr>
<th>Regression model</th>
<th>Kernel</th>
</tr>
</thead>
<tbody>
<tr>
<td>GP smoothing</td>
<td>SE + WN</td>
</tr>
<tr>
<td>Linear regression</td>
<td>C + LIN + WN</td>
</tr>
<tr>
<td>Multiple kernel learning</td>
<td>$\sum$ SE + WN</td>
</tr>
<tr>
<td>Trend, cyclical, irregular</td>
<td>$\sum$ SE + $\sum$ PER + WN</td>
</tr>
<tr>
<td>Fourier decomposition</td>
<td>C + $\sum$ cos + WN</td>
</tr>
<tr>
<td>Sparse spectrum GPs</td>
<td>$\sum$ cos + WN</td>
</tr>
<tr>
<td>Spectral mixture</td>
<td>$\sum$ SE $\times$ cos + WN</td>
</tr>
<tr>
<td>Changepoints</td>
<td>e.g. CP(SE, SE) + WN</td>
</tr>
<tr>
<td>Heteroscedasticity</td>
<td>e.g. SE + LIN $\times$ WN</td>
</tr>
</tbody>
</table>

Note: cos is a special case of our version of PER
DISCOVERING A GOOD MODEL VIA SEARCH

Language of models

Data -> Search

Evaluation

Model -> Prediction

Checking

Report

Translation
Language defined as the arbitrary composition of five base kernels (WN, C, LIN, SE, PER) via three operators (+, ×, CP).

The space spanned by this language is open-ended and can have a high branching factor requiring a judicious search.

We propose a greedy search for its simplicity and similarity to human model-building.
EXAMPLE: MAUNA LOA KEELING CURVE

![Graph showing the Mauna Loa Keeling Curve with an example equation: SE × (PER + RQ) and other nodes labeled Start, RQ, LIN, PER.](image-url)
**Example: Mauna Loa Keeling Curve**

![Graph showing the Mauna Loa Keeling Curve with a trend line and labels for PER + RQ, SE + RQ, PER x RQ, and other nodes in a tree diagram.](image-url)

The graph illustrates the increase in atmospheric CO₂ concentration from 2000 to 2010, highlighting the PER + RQ component. The tree diagram to the right elaborates on the relationships between SE, RQ, LIN, and PER, with PER + RQ being a key node.
**Example: Mauna Loa Keeling Curve**

\[ SE \times (\text{Per} + \text{RQ}) \]

- \( SE \times (\text{Per} + \text{RQ}) \)
- \( \text{SE} \)
- \( \text{RQ} \)
- \( \text{LIN} \)
- \( \text{PER} \)
- \( \text{SE} + \text{RQ} \)
- \( \text{PER} + \text{RQ} \)
- \( \text{SE} \times (\text{PER} + \text{RQ}) \)
**Example: Mauna Loa Keeling Curve**

\[(SE + SE \times (Per + RQ))\]

Graph showing a line plot with years 2000, 2005, and 2010 on the x-axis and measurements on the y-axis.

Diagram with nodes labeled as:
- Start
- \(SE\)
- \(RQ\)
- \(LIN\)
- \(PER\)
- \(SE + RQ\)
- \(PER + RQ\)
- \(SE \times (PER + RQ)\)
- \(SE + PER + RQ\)
- \(PER \times RQ\)

James Robert Lloyd and Zoubin Ghahramani
MODEL EVALUATION

Language of models

Data → Search → Evaluation → Model → Prediction → Report

Translation

Checking
After proposing a new model its kernel parameters are optimised by conjugate gradients.

We evaluate each optimised model, $M$, using the marginal likelihood which can be computed analytically for GPs.

We penalise the marginal likelihood for the optimised kernel parameters using the Bayesian Information Criterion (BIC):

$$-0.5 \times \text{BIC}(M) = \log p(D \mid M) - \frac{p}{2} \log n$$

where $p$ is the number of kernel parameters, $D$ represents the data, and $n$ is the number of data points.
AUTOMATIC TRANSLATION OF MODELS

Language of models

Data → Search → Model → Prediction → Report → Checking → Evaluation

Translation
Search can produce arbitrarily complicated models from open-ended language but two main properties allow description to be automated:

- Kernels can be decomposed into a sum of products:
  - A sum of kernels corresponds to a sum of functions
  - Therefore, we can describe each product of kernels separately

- Each kernel in a product modifies a model in a consistent way:
  - Each kernel roughly corresponds to an adjective
Suppose the search finds the following kernel

\[ SE \times (WN \times LIN + CP(C, PER)) \]
Suppose the search finds the following kernel

$$SE \times (WN \times LIN + CP(C, PER))$$

The changepoint can be converted into a sum of products

$$SE \times (WN \times LIN + C \times \sigma + PER \times \bar{\sigma})$$
Suppose the search finds the following kernel

\[ SE \times (WN \times LIN + CP(C, PER)) \]

The changepoint can be converted into a sum of products

\[ SE \times (WN \times LIN + C \times \sigma + PER \times \bar{\sigma}) \]

Multiplication can be distributed over addition

\[ SE \times WN \times LIN + SE \times C \times \sigma + SE \times PER \times \bar{\sigma} \]
Suppose the search finds the following kernel

\[ SE \times (WN \times LIN + CP(C, PER)) \]

The changepoint can be converted into a sum of products

\[ SE \times (WN \times LIN + C \times \sigma + PER \times \bar{\sigma}) \]

Multiplication can be distributed over addition

\[ SE \times WN \times LIN + SE \times C \times \sigma + SE \times PER \times \bar{\sigma} \]

Simplification rules are applied

\[ WN \times LIN + SE \times \sigma + SE \times PER \times \bar{\sigma} \]
**Sums of Kernels Are Sums of Functions**

If \( f_1 \sim \text{GP}(0, k_1) \) and independently \( f_2 \sim \text{GP}(0, k_2) \) then

\[
f_1 + f_2 \sim \text{GP}(0, k_1 + k_2)
\]

e.g.

We can therefore describe each component separately.
PRODUCTS OF KERNELS

PER

periodic function

On their own, each kernel is described by a standard noun phrase

---

---

---

---
PRODUCTS OF KERNELS - SE

\[
\text{SE} \times \text{PER}
\]

approximately periodic function

**Multiplication by SE** removes long range correlations from a model since \( \text{SE}(x, x') \) decreases monotonically to 0 as \( |x - x'| \) increases.
**PRODUCTS OF KERNELS - LIN**

\[
\text{SE} \times \text{PER} \times \text{LIN}
\]

approximately periodic function with linearly growing amplitude

**Multiplication by LIN** is equivalent to multiplying the function being modeled by a linear function. If \( f(x) \sim \text{GP}(0, k) \), then \( xf(x) \sim \text{GP}(0, k \times \text{LIN}) \). This causes the standard deviation of the model to vary linearly without affecting the correlation.
PRODUCTS OF KERNELS - CHANGEPOINTS

\[ \text{SE} \times \text{PER} \times \text{LIN} \times \sigma \]

approximately periodic function with linearly growing amplitude until 1700

**Multiplication by** $\sigma$ **is equivalent to multiplying the function being modeled by a sigmoid.**
AUTOMATICALLY GENERATED REPORTS
**Example: Airline Passenger Volume**

Four additive components have been identified in the data:

- A linearly increasing function.
- An approximately periodic function with a period of 1.0 years and with linearly increasing amplitude.
- A smooth function.
- Uncorrelated noise with linearly increasing standard deviation.
EXAMPLE: AIRLINE PASSENGER VOLUME

This component is linearly increasing.
This component is approximately periodic with a period of 1.0 years and varying amplitude. Across periods the shape of this function varies very smoothly. The amplitude of the function increases linearly. The shape of this function within each period has a typical lengthscale of 6.0 weeks.
**Example: Airline Passenger Volume**

This component is a smooth function with a typical lengthscale of 8.1 months.
This component models uncorrelated noise. The standard deviation of the noise increases linearly.
Example: Solar Irradiance

This component is constant.
**Example: Solar Irradiance**

This component is constant. This component applies from 1643 until 1716.
EXAMPLE: SOLAR IRRADIANCE

This component is a smooth function with a typical lengthscale of 23.1 years. This component applies until 1643 and from 1716 onwards.
EXAMPLE: SOLAR IRRADIANCE

This component is approximately periodic with a period of 10.8 years. Across periods the shape of this function varies smoothly with a typical lengthscale of 36.9 years. The shape of this function within each period is very smooth and resembles a sinusoid. This component applies until 1643 and from 1716 onwards.
EXAMPLE: CALL CENTRE VOLUME

See pdf
Good predictive performance as well

Standardised RMSE over 13 data sets

- Tweaks can be made to the algorithm to improve accuracy or interpretability of models produced...

- ...but both methods are highly competitive at extrapolation (shown above) and interpolation
CHALLENGES

- Interpretability / accuracy

- Increasing the expressivity of language
  - e.g. Monotonocity, positive functions, symmetries

- Computational complexity of searching through a huge space of models

- Extending the automatic reports to multidimensional datasets
  - Search and descriptions naturally extend to multiple dimensions, but automatically generating relevant visual summaries harder
CURRENT AND FUTURE DIRECTIONS

- Automatic statistician for:
  - Multivariate nonlinear regression
  - Classification
  - Completing and interpreting tables and databases

- Probabilistic programming
  - Probabilistic models are expressed in a general (Turing complete) programming language (e.g. Church/Venture/Anglican)
  - A universal inference engine can then be used to infer unobserved variables given observed data
  - This can be used to implement search over the model space in an automated statistician
SUMMARY

- We have presented the beginnings of an automatic statistician for regression

- Our system
  - Defines an open-ended language of models
  - Searches greedily through this space
  - Produces detailed reports describing patterns in data

- Extrapolation and interpolation performance highly competitive

- We believe this line of research has the potential to make powerful statistical model-building techniques accessible to non-experts
Visit the website

www.automaticstatistician.com