The Automatic Statistician and Future Directions in Probabilistic Machine Learning

Zoubin Ghahramani

Department of Engineering
University of Cambridge

zoubin@eng.cam.ac.uk
http://mlg.eng.cam.ac.uk/
http://www.automaticstatistician.com/

MLSS 2015, Tübingen

MACHINE LEARNING AS PROBABILISTIC MODELLING

- A model describes data that one could observe from a system
- If we use the mathematics of probability theory to express all forms of uncertainty and noise associated with our model...
- ...then *inverse probability* (i.e. Bayes rule) allows us to infer unknown quantities, adapt our models, make predictions and learn from data.

Bayes Rule

$$
\begin{aligned}
P(\text{hypothesis}|\text{data}) &= \frac{P(\text{data}|\text{hypothesis})P(\text{hypothesis})}{P(\text{data})} \\
&= \frac{P(\text{data}|\text{hypothesis})P(\text{hypothesis})}{\sum_{\text{h}} P(\text{data}|\text{h})P(\text{h})}
\end{aligned}
$$

BAYESIAN MACHINE LEARNING

Everything follows from two simple rules:

Sum rule: $P(x) = \sum_y P(x,y)$

Product rule: $P(x,y) = P(x)P(y|x)$

Learning:

$$P(\theta|\mathcal{D},m) = \frac{P(\mathcal{D}|\theta,m)P(\theta|m)}{P(\mathcal{D}|m)}$$

$P(\mathcal{D}\|\theta,m)$	likelihood of parameters θ in model m
$P(\theta\|m)$	prior probability of θ
$P(\theta\|\mathcal{D},m)$	posterior of θ given data $\mathcal{D}$

Prediction:

$$P(x|\mathcal{D},m) = \int P(x|\theta,\mathcal{D},m)P(\theta|\mathcal{D},m)d\theta$$

Model Comparison:

$$P(m|\mathcal{D}) = \frac{P(\mathcal{D}|m)P(m)}{P(\mathcal{D})}$$

When is the probabilistic approach essential?

Many aspects of learning and intelligence depend crucially on the careful probabilistic representation of *uncertainty*:

- Forecasting
- Decision making
- Learning from limited, noisy, and missing data
- Learning complex personalised models
- Data compression
- Automating scientific modelling, discovery, and experiment design

CURRENT AND FUTURE DIRECTIONS

- Probabilistic programming
- Bayesian optimisation
- Rational allocation of computational resources
- Probabilistic models for efficient data compression
- The automatic statistician

PROBABILISTIC PROGRAMMING

Problem: Probabilistic model development and the derivation of inference algorithms is time-consuming and error-prone.

Probabilistic Programming

Problem: Probabilistic model development and the derivation of inference algorithms is time-consuming and error-prone.
Solution:

- Develop Turing-complete **Probabilistic Programming Languages** for expressing probabilistic models as computer programs that generate data (i.e. simulators).
- Derive **Universal Inference Engines** for these languages that sample over program traces given observed data.

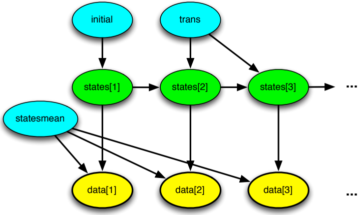
Example languages: Church, Venture, Anglican, Stochastic Python*, ones based on Haskell*, Julia*

Example inference algorithms: Metropolis-Hastings MCMC, variational inference, particle filtering, slice sampling*, particle MCMC, nested particle inference*, austerity MCMC*

Probabilistic Programming

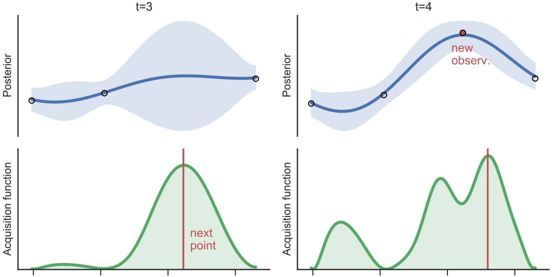
```
statesmean = [-1, 1, 0]  # Emission parameters.
initial    = Categorical([1.0/3, 1.0/3, 1.0/3]) # Prob distr of state[1].
trans      = [Categorical([0.1, 0.5, 0.4]), Categorical([0.2, 0.2, 0.6]),
              Categorical([0.15, 0.15, 0.7])]   # Trans distr for each state.
data       = [Nil, 0.9, 0.8, 0.7, 0, -0.025, -5, -2, -0.1, 0, 0.13]

@model hmm begin # Define a model hmm.
 states = Array(Int, length(data))
 @assume(states[1] ~ initial)
 for i = 2:length(data)
   @assume(states[i] ~ trans[states[i-1]])
   @observe(data[i]  ~ Normal(statesmean[states[i]], 0.4))
 end
 @predict states
end
```

An example probabilistic program in Julia implementing a 3-state hidden Markov model (HMM).



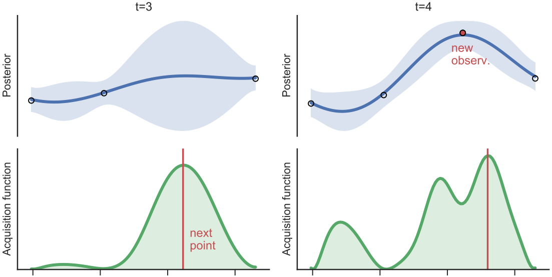
Probabilistic programming could revolutionise scientific modelling.

Bayesian Optimisation



Problem: Global optimisation of black-box functions that are *expensive to evaluate*

Bayesian Optimisation



Problem: Global optimisation of black-box functions that are *expensive to evaluate*

Solution: treat as a problem of sequential decision-making and model uncertainty in the function.

This has myriad applications, from robotics to drug design, to learning neural networks, and speeding up model search in the automatic statistician.

Bayesian Optimisation

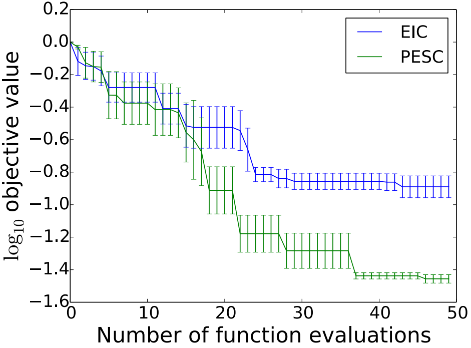


Figure 4. Classification error of a 3-hidden-layer neural network constrained to make predictions in under 2 ms.

(work with J.M. Hernández-Lobato, M.A. Gelbart, M.W. Hoffman, & R.P. Adams)

Rational allocation of computational resources

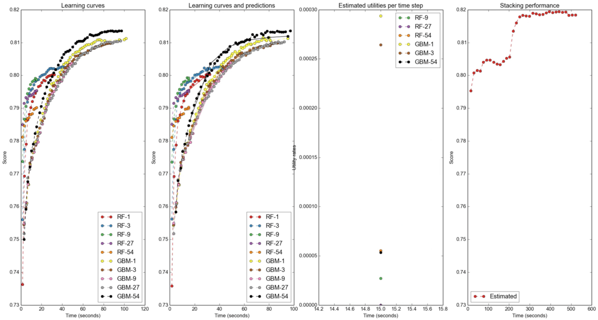
Problem: Many problems in machine learning and AI require the evaluation of a large number of alternative models on potentially large datasets. A rational agent needs to consider the *tradeoff between statistical and computational efficiency.*

RATIONAL ALLOCATION OF COMPUTATIONAL RESOURCES

Problem: Many problems in machine learning and AI require the evaluation of a large number of alternative models on potentially large datasets. A rational agent needs to consider the *tradeoff between statistical and computational efficiency*.

Solution: Treat the allocation of computational resources as a problem in sequential decision-making under uncertainty.

Rational allocation of computational resources



Movie Link

(work with James R. Lloyd)

PROBABILISTIC DATA COMPRESSION

Problem: We often produce more data than we can store or transmit.
(E.g. CERN → data centres, or Mars Rover → Earth.)

PROBABILISTIC DATA COMPRESSION

Problem: We often produce more data than we can store or transmit. (E.g. CERN → data centres, or Mars Rover → Earth.)

Solution:

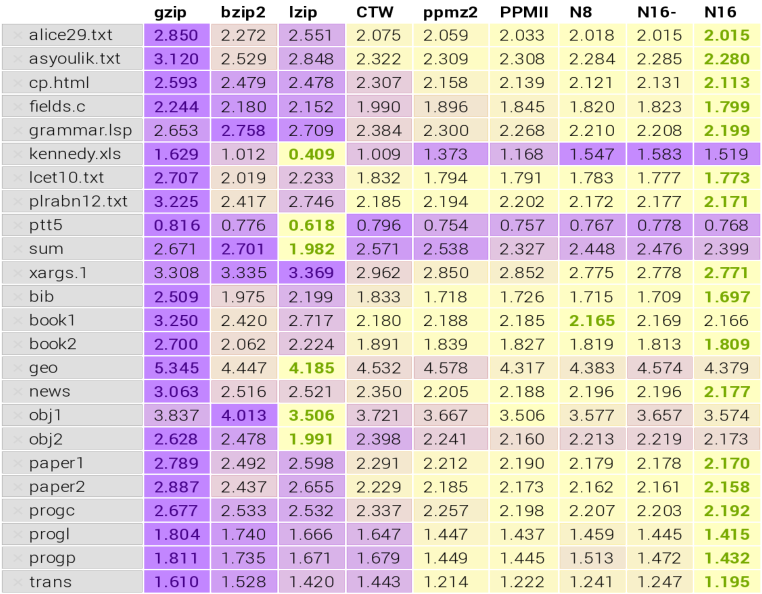
- Use the same resources more effectively by *predicting the data* with a probabilistic model.
- Produce a description of the data that is (on average) cheaper to store or transmit.

Example: "PPM-DP" is based on a probabilistic model that learns and predicts symbol occurences in sequences. It works on arbitrary files, but delivers cutting-edge compression results for human text.

Probabilistic models for human text also have many other applications aside from data compression, e.g. smart text entry methods, anomaly detection, sequence synthesis.

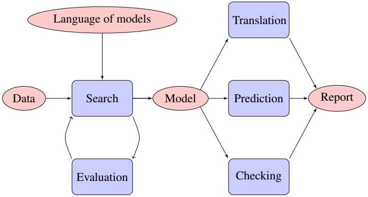
(work with Christian Steinruecken and David J. C. MacKay)

PROBABILISTIC DATA COMPRESSION



	gzip	bzip2	lzip	CTW	ppmz2	PPMII	N8	N16-	N16
alice29.txt	2.850	2.272	2.551	2.075	2.059	2.033	2.018	2.015	2.015
asyoulik.txt	3.120	2.529	2.848	2.322	2.309	2.308	2.284	2.285	2.280
cp.html	2.593	2.479	2.478	2.307	2.158	2.139	2.121	2.131	2.113
fields.c	2.244	2.180	2.152	1.990	1.896	1.845	1.820	1.823	1.799
grammar.lsp	2.653	2.758	2.709	2.384	2.300	2.268	2.210	2.208	2.199
kennedy.xls	1.629	1.012	0.409	1.009	1.373	1.168	1.547	1.583	1.519
lcet10.txt	2.707	2.019	2.233	1.832	1.794	1.791	1.783	1.777	1.773
plrabn12.txt	3.225	2.417	2.746	2.185	2.194	2.202	2.172	2.177	2.171
ptt5	0.816	0.776	0.618	0.796	0.754	0.757	0.767	0.778	0.768
sum	2.671	2.701	1.982	2.571	2.538	2.327	2.448	2.476	2.399
xargs.1	3.308	3.335	3.369	2.962	2.850	2.852	2.775	2.778	2.771
bib	2.509	1.975	2.199	1.833	1.718	1.726	1.715	1.709	1.697
book1	3.250	2.420	2.717	2.180	2.188	2.185	2.165	2.169	2.166
book2	2.700	2.062	2.224	1.891	1.839	1.827	1.819	1.813	1.809
geo	5.345	4.447	4.185	4.532	4.578	4.317	4.383	4.574	4.379
news	3.063	2.516	2.521	2.350	2.205	2.188	2.196	2.196	2.177
obj1	3.837	4.013	3.506	3.721	3.667	3.506	3.577	3.657	3.574
obj2	2.628	2.478	1.991	2.398	2.241	2.160	2.213	2.219	2.173
paper1	2.789	2.492	2.598	2.291	2.212	2.190	2.179	2.178	2.170
paper2	2.887	2.437	2.655	2.229	2.185	2.173	2.162	2.161	2.158
progc	2.677	2.533	2.532	2.337	2.257	2.198	2.207	2.203	2.192
progl	1.804	1.740	1.666	1.647	1.447	1.437	1.459	1.445	1.415
progp	1.811	1.735	1.671	1.679	1.449	1.445	1.513	1.472	1.432
trans	1.610	1.528	1.420	1.443	1.214	1.222	1.241	1.247	1.195

The Automatic Statistician

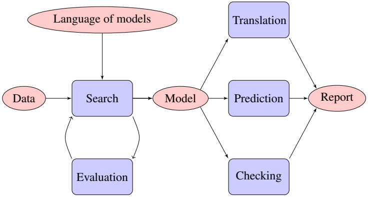


Problem: Data are now ubiquitous; there is great value from understanding this data, building models and making predictions... however, *there aren't enough data scientists, statisticians, and machine learning experts.*

Solution: Develop a system that automates model discovery from data:

- processing data, searching over models, discovering a good model, and explaining what has been discovered to the user.

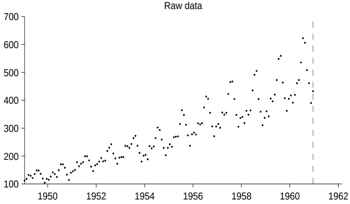
THE AUTOMATIC STATISTICIAN

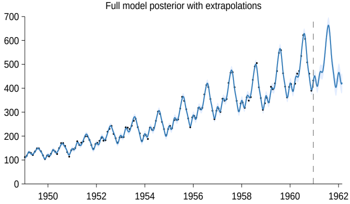


- **An open-ended language of models**
  - Expressive enough to capture real-world phenomena…
  - …and the techniques used by human statisticians
- **A search procedure**
  - To efficiently explore the language of models
- **A principled method of evaluating models**
  - Trading off complexity and fit 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

(work with J. R. Lloyd, D.Duvenaud, R.Grosse, and J.B.Tenenbaum)

Example: An entirely automatic analysis

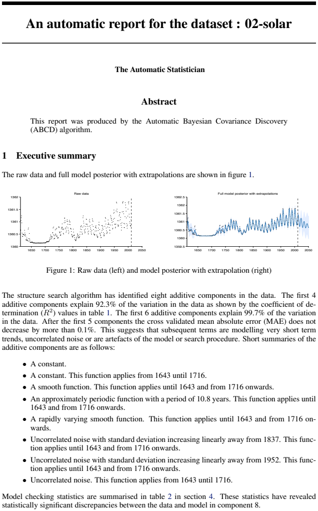




Four additive components have been identified in the data

- 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 Reports



An automatic report for the dataset : 02-solar

The Automatic Statistician

Abstract

This report was produced by the Automatic Bayesian Covariance Discovery (ABCD) algorithm.

1 Executive summary

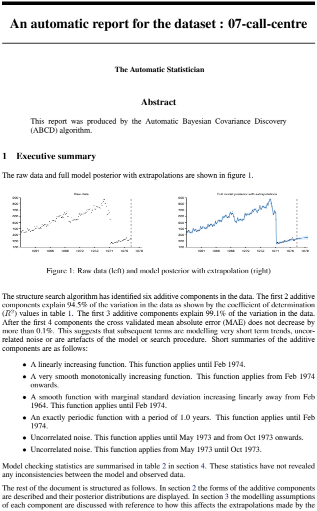
The raw data and full model posterior with extrapolations are shown in figure 1.

Figure 1: Raw data (left) and model posterior with extrapolation (right)

The structure search algorithm has identified eight additive components in the data. The first 4 additive components explain 92.3% of the variation in the data as shown by the coefficient of determination (R^2) values in table 1. The first 6 additive components explain 99.7% of the variation in the data. After the first 5 components the cross validated mean absolute error (MAE) does not decrease by more than 0.1%. This suggests that subsequent terms are modelling very short term trends, uncorrelated noise or are artefacts of the model or search procedure. Short summaries of the additive components are as follows:

- A constant.
- A constant. This function applies from 1643 until 1716.
- A smooth function. This function applies until 1643 and from 1716 onwards.
- An approximately periodic function with a period of 10.8 years. This function applies until 1643 and from 1716 onwards.
- A rapidly varying smooth function. This function applies until 1643 and from 1716 onwards.
- Uncorrelated noise with standard deviation increasing linearly away from 1837. This function applies until 1643 and from 1716 onwards.
- Uncorrelated noise with standard deviation increasing linearly away from 1952. This function applies until 1643 and from 1716 onwards.
- Uncorrelated noise. This function applies from 1643 until 1716.

Model checking statistics are summarised in table 2 in section 4. These statistics have revealed statistically significant discrepancies between the data and model in component 8.



An automatic report for the dataset : 07-call-centre

The Automatic Statistician

Abstract

This report was produced by the Automatic Bayesian Covariance Discovery (ABCD) algorithm.

1 Executive summary

The raw data and full model posterior with extrapolations are shown in figure 1.

Figure 1: Raw data (left) and model posterior with extrapolation (right)

The structure search algorithm has identified six additive components in the data. The first 2 additive components explain 94.5% of the variation in the data as shown by the coefficient of determination (R^2) values in table 1. The first 3 additive components explain 99.1% of the variation in the data. After the first 4 components the cross validated mean absolute error (MAE) does not decrease by more than 0.1%. This suggests that subsequent terms are modelling very short term trends, uncorrelated noise or are artefacts of the model or search procedure. Short summaries of the additive components are as follows:

- A linearly increasing function. This function applies until Feb 1974.
- A very smooth monotonically increasing function. This function applies from Feb 1974 onwards.
- A smooth function with marginal standard deviation increasing linearly away from Feb 1964. This function applies until Feb 1974.
- An exactly periodic function with a period of 1.0 years. This function applies until Feb 1974.
- Uncorrelated noise. This function applies until May 1973 and from Oct 1973 onwards.
- Uncorrelated noise. This function applies from May 1973 until Oct 1973.

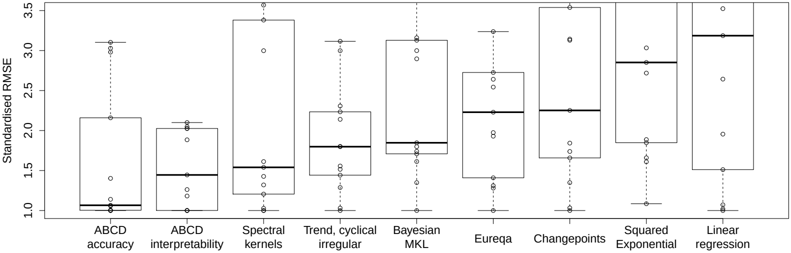
Model checking statistics are summarised in table 2 in section 4. These statistics have not revealed any inconsistencies between the model and observed data.

The rest of the document is structured as follows. In section 2 the forms of the additive components are described and their posterior distributions are displayed. In section 3 the modelling assumptions of each component are discussed with reference to how this affects the extrapolations made by the

See http://www.automaticstatistician.com

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

SUMMARY: THE AUTOMATIC STATISTICIAN

- We have presented the beginnings of an automatic statistician

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

- 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

CONCLUSIONS

Probabilistic modelling offers a framework for building systems that reason about uncertainty and learn from data, going beyond traditional pattern recognition problems.

I have reviewed some of the frontiers of research, including:

- Probabilistic programming
- Bayesian optimisation
- Rational allocation of computational resources
- Probabilistic models for efficient data compression
- The automatic statistician

Thanks!

Appendix: Model Checking and Criticism

- Good statistical modelling should include model criticism:
  - Does the data match the assumptions of the model?
  - For example, if the model assumed Gaussian noise, does a Q-Q plot reveal non-Gaussian residuals?
- Our automatic statistician does posterior predictive checks, dependence tests and residual tests
- We have also been developing more systematic nonparametric approaches to model criticism using kernel two-sample testing with MMD.

Lloyd, J. R., and Ghahramani, Z. (2014) Statistical Model Criticism using Kernel Two Sample Tests. http://mlg.eng.cam.ac.uk/Lloyd/papers/kernel-model-checking.pdf

PAPERS

General:

Ghahramani, Z. (2013) Bayesian nonparametrics and the probabilistic approach to modelling. *Philosophical Trans. Royal Society A* **371**: 20110553.

Ghahramani, Z. (2015) Probabilistic machine learning and artificial intelligence *Nature* **521**:452–459. http://www.nature.com/nature/journal/v521/n7553/full/nature14541.html

Automatic Statistician:

Website: http://www.automaticstatistician.com

Duvenaud, D., Lloyd, J. R., Grosse, R., Tenenbaum, J. B. and Ghahramani, Z. (2013) Structure Discovery in Nonparametric Regression through Compositional Kernel Search. ICML 2013.

Lloyd, J. R., Duvenaud, D., Grosse, R., Tenenbaum, J. B. and Ghahramani, Z. (2014) Automatic Construction and Natural-language Description of Nonparametric Regression Models AAAI 2014. http://arxiv.org/pdf/1402.4304v2.pdf

Lloyd, J. R., and Ghahramani, Z. (2014) Statistical Model Criticism using Kernel Two Sample Tests http://mlg.eng.cam.ac.uk/Lloyd/papers/kernel-model-checking.pdf

Papers II

Bayesian Optimisation:

Hernández-Lobato, J. M., Hoffman, M. W., and Ghahramani, Z. (2014) Predictive entropy search for efficient global optimization of black-box functions. NIPS 2014

Hernández-Lobato, J.-M. Gelbart, M. A., Hoffman, M. W., Adams, R. P., Ghahramani, Z. (2015) Predictive Entropy Search for Bayesian Optimization with Unknown Constraints. arXiv:1502.05312

Data Compression:

Steinruecken, C., Ghahramani, Z. and MacKay, D.J.C. (2015) Improving PPM with dynamic parameter updates. Data Compression Conference (DCC 2015). Snowbird, Utah.

Probabilistic Programming:

Chen, Y., Mansinghka, V., Ghahramani, Z. (2014) Sublinear-Time Approximate MCMC Transitions for Probabilistic Programs. arXiv:1411.1690