GEFCom2012 Hierarchical load forecasting: Gradient boosting machines and Gaussian processes

James Robert Lloyd

Machine Learning Group,
Department of Engineering,
University of Cambridge

July 2013

Thanks to
Alex Davies
David Duvenaud
Zoubin Ghahramani

Overview of techniques

Preprocessing

- Kernel smoothing of temperatures (to remove daily periodicity)

Temperature forecasting

- Gaussian process (GP) regression

Load back/forecasting

- Gradient boosting machine (GBM) regression - 76%
- Gaussian process (GP) regression - 14%
- Linear regression (benchmark solution) - 10%

Performance of Different Components

Method	Validation score
GBM	72,968
GP	99,787
LR	112,547
Ensemble	71,164

- GBM the best performing method
- GP and LR sufficiently uncorrelated with GBM to provide useful components in ensemble

Main approach - regression on time and temp.

Modelling temperatures and loads as functions of time and temperature

$$
\begin{aligned}
T(t) &= f(t, \bar{S}(t)) + \varepsilon_t^T \\
Z(t) &= g(t, T(t), S(t)) + \varepsilon_t^Z
\end{aligned}
$$

i.e. no explicit autoregressive components e.g.

$$
y(t+1) = f(y(t)) + \varepsilon_t
$$

Notation

t — Time, T — Temperature, S — Smoothed temperature $\bar{S}$ — Historical average of smoothed temperature, Z — Load, f, g — Generic functions, ε — Generic error

GBM REGRESSION - OVERVIEW

Used as a regression 'black-box'

- Bagged and boosted decision trees
- Used standard R implementation with most default parameters unchanged

Output

- Z_i i.e. each load zone modelled in isolation

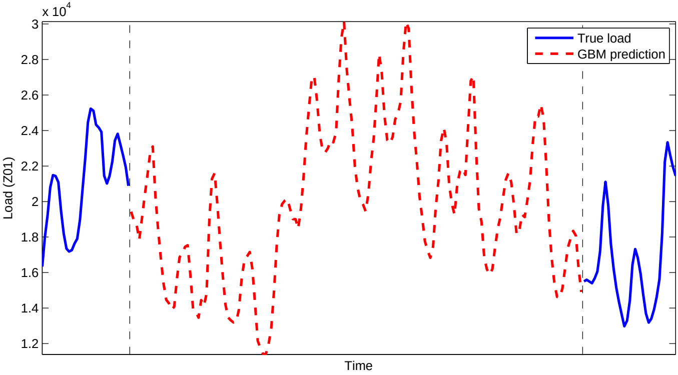
Inputs

- Time of day
- Time within week
- Temperatures (all stations)
- Smoothed temperatures (all stations)

GBM REGRESSION - PARAMETER SELECTION

- Ideally would have performed grid searches over parameter values using cross validated error as metric
- In practice, partial grid searches combined with intuition, using out of bag errors and validation score on Kaggle
- 10,000 trees, interaction depth of 3 and shrinkage factor of 0.01 (other values set to defaults of R implementation)

GBM REGRESSION - EXAMPLE



Only slight discontinuity between prediction and ground truth despite no explicit modelling assumptions of continuity

Gaussian process regression

- A Bayesian nonparametric method for regression
- Places a prior on functions but equivalent to linear regression in an (infinite dimensional) feature space
- Typically used as a smoothing device by choosing a default *kernel*
- Data exhibiting high level structure (e.g. periodicity) can be modelled using more advanced kernels

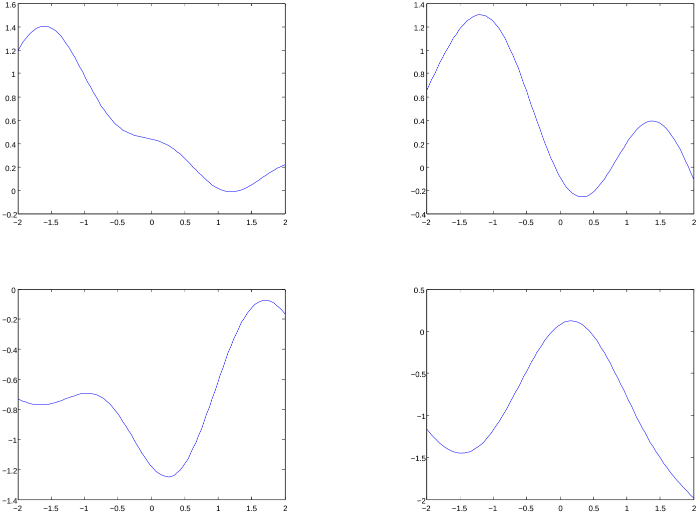
Bayesian modelling

Bayes' rule

$$\mathbb{P}(\text{hypothesis}|\text{data}) = \frac{\mathbb{P}(\text{data}|\text{hypothesis})\mathbb{P}(\text{hypothesis})}{\mathbb{P}(\text{data})}$$

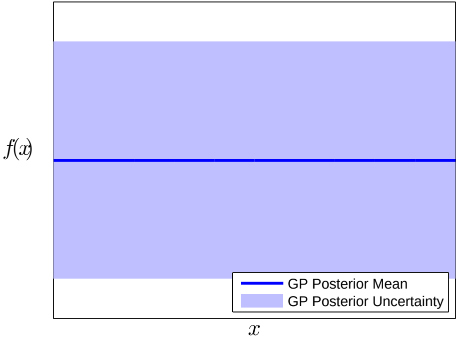
- Bayes' rule follows from basic axioms of probability theory
- Provides a calculus to update beliefs in response to data
- Requires the specification of prior beliefs about data - choice of prior is crucial for successful modelling

PRIOR ON FUNCTIONS



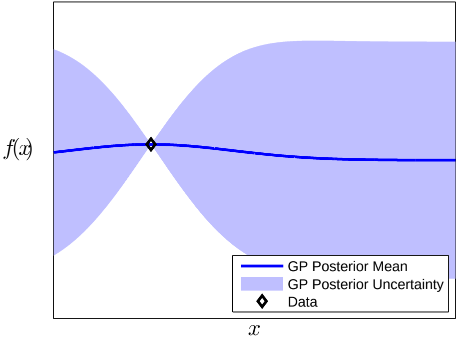
Conditional Posterior

After observing data, Bayes rule provides a formula by which to update our beliefs about a function



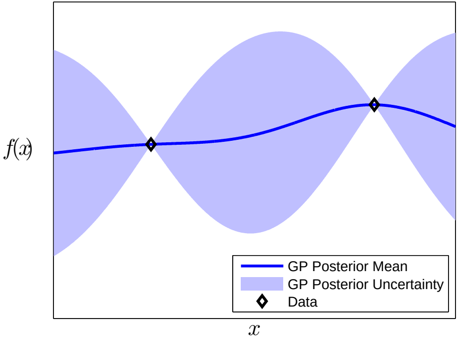
Conditional Posterior

After observing data, Bayes rule provides a formula by which to update our beliefs about a function



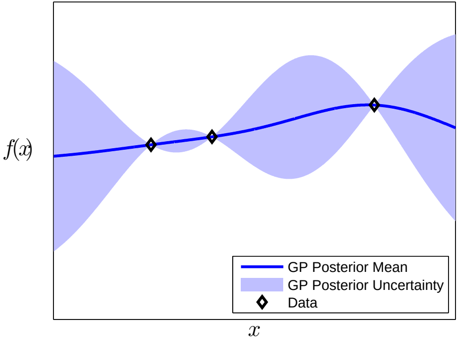
Conditional Posterior

After observing data, Bayes rule provides a formula by which to update our beliefs about a function



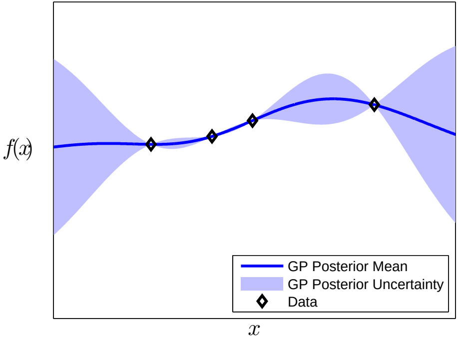
CONDITIONAL POSTERIOR

After observing data, Bayes rule provides a formula by which to update our beliefs about a function



CONDITIONAL POSTERIOR

After observing data, Bayes rule provides a formula by which to update our beliefs about a function

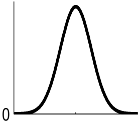


ENCODING STRUCTURAL ASSUMPTIONS

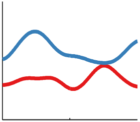
- Gaussian processes typically used as smoothing devices
- Daily and weekly periodicity assumptions could be encoded by feature engineering as with GBM
- However, structural assumptions can also be encoded in the *kernel* of a GP
  - Using a different method allows predictions to be uncorrelated - very useful for ensembling

CAN ENCODE STRUCTURAL ASSUMPTIONS IN KERNEL

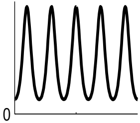
- Kernel determines the structural properties of a Gaussian process
- Many different kinds, with very different properties:



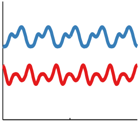
Squared-exp (SE)



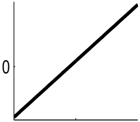
smooth functions



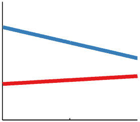
Periodic (PER)



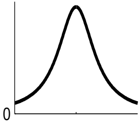
repeating structure



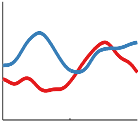
Linear (LIN)



linear functions



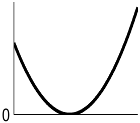
Rational-quadratic(RQ)



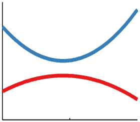
multi-scale variation

Kernels can be composed

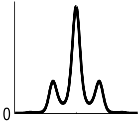
- Two main operations: adding, multiplying



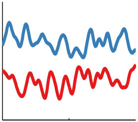
Lin $\times$ Lin



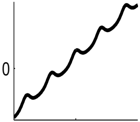
quadratic functions



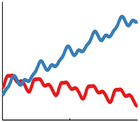
SE $\times$ Per



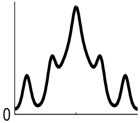
locally periodic



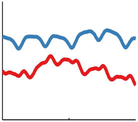
Lin $+$ Per



periodic with trend



SE $+$ Per

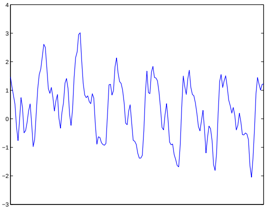


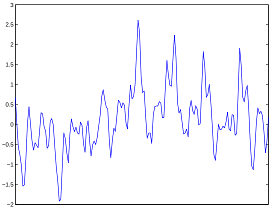
periodic with noise

A SUITABLE PRIOR

Used structured kernel to encode the assumption that

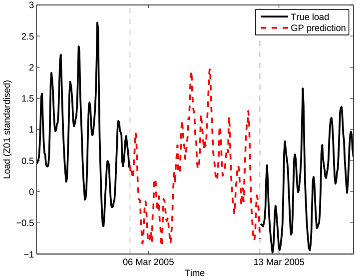
$$\begin{aligned} \text{Load} \quad &= \quad \text{Smooth function of time} + \\ &\quad\;\; \text{Smooth function of smoothed temperatures} + \\ &\quad\;\; \text{Daily periodicity smoothly changing with time and temperature} \end{aligned}$$

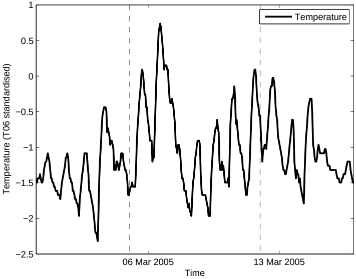




GP REGRESSION - EXAMPLE

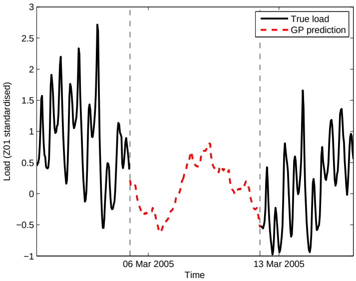
Load = Smooth function of time +
Smooth function of smoothed temperatures +
Daily periodicity smoothly changing with time and temperature

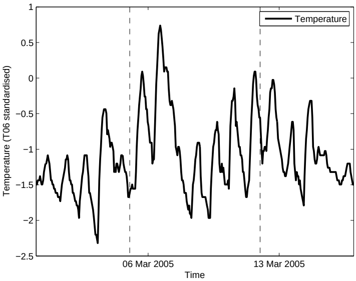




INCORRECT PRIOR - NO PERIODICITY

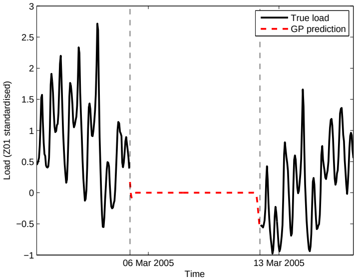
$$
\begin{aligned}
\text{Load} \quad = \quad & \text{Smooth function of time} + \\
& \text{Smooth function of temperatures} + \\
& \text{Smooth function of smoothed temperatures}
\end{aligned}
$$

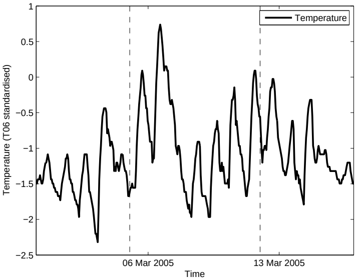




INCORRECT PRIOR - NO STRUCTURE

$$\text{Load} \quad = \quad \text{Smooth function of time}$$

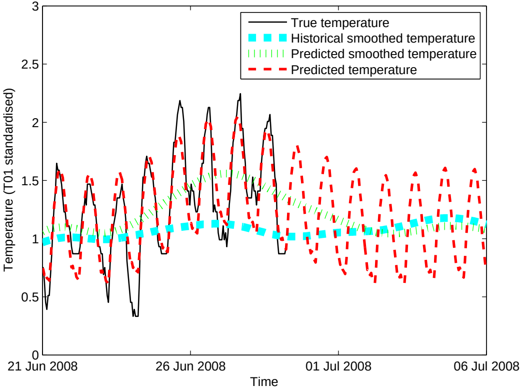




Suitable prior assumptions crucial when using any Bayesian method

STRUCTURED KERNELS FOR TEMP. FORECASTS

$$\begin{aligned} \text{Temperature} \quad &= \quad \text{Smooth historical average temperature} + \\ &\quad\ \text{Smooth long-term deviations} + \\ &\quad\ \text{Daily periodicity} \end{aligned}$$



GP REGRESSION - PARAMETER SELECTION

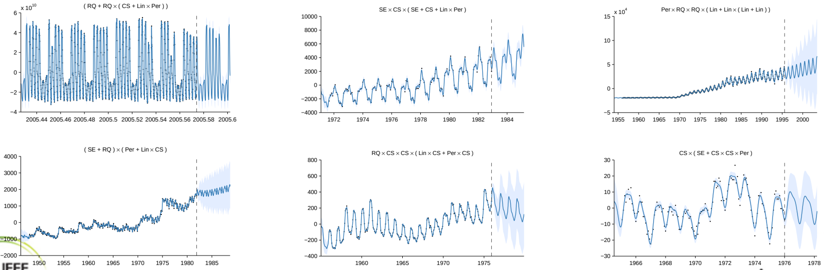
- Can optimise marginal likelihood (a balance of model fit and complexity) with gradient based optimisation

- Marginal likelihood optimisation can fail
  - Can result in slight over fitting
  - When the prior and data generation process are dissimilar, Bayesian inference can give misleading results

- In practice, parameter selection was a mixture of marginal likelihood optimisation, validation score maximisation and model checking (plotting graphs)

Competition inspired new GP research

- Creating custom composite kernels not a new idea, but typically only practised by GP / kernel learning experts
- After competition, automated the process of kernel / model construction [DLG+13] based on an idea by [GSFT12] in the context of matrix factorisation
- Ongoing research to see how far the automatic model construction idea can be pushed e.g.



ENSEMBLING

Small search over possible weightings

GBM	GP	RF	LM	Score
100	0	0	0	72,968
0	100	0	0	99,787
0	0	100	0	89,457
0	0	0	100	112,547
80	20	0	0	71,683
70	30	0	0	72,485
90	10	0	0	71,846
85	15	0	0	71,644
76	**14**	**0**	**10**	**71,164**
72	13	10	5	71,566
80	0	20	10	74,293
...	...	...	...	...

More principled methods

- Grid searches and cross validation - but could be costly to retrain algorithms on different training / test splits
- Bayesian optimisation [OGR09], [SLA12], [HS12] can be more appropriate when individual evaluations costly (e.g. submitting to Kaggle to obtain validation score)

SUMMARY

- Main approach was to regress loads on time and temperature, rather than using an autoregressive model
- GBM provided the majority of performance
- Structured kernel GP method sufficiently uncorrelated to provide useful component in ensemble

REFERENCES I

[DLG+13] D. Duvenaud, J. R. Lloyd, R. Grosse, J. B. Tenenbaum, and Z. Ghahramani. Structure discovery in nonparametric regression through compositional kernel search. In *Proceedings of the 30th International Conference on Machine Learning*, 2013.

[GSFT12] Roger B. Grosse, Ruslan Salakhutdinov, William T. Freeman, and Joshua B. Tenenbaum. Exploiting compositionality to explore a large space of model structures. In *Conference on Uncertainty in Artificial Intelligence*, 2012.

[HS12] Philipp Hennig and Christian J. Schuler. Entropy search for information-efficient global optimization. *Journal of Machine Learning Research*, 13, 2012.

[OGR09] Michael A. Osborne, Roman Garnett, and Stephen J. Roberts. Gaussian processes for global optimization. In *3rd International Conference on Learning and Intelligent Optimization (LION3)*, 2009.

[SLA12] Jasper Snoek, Hugo Larochelle, and Ryan P. Adams. Practical bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems*, 2012.