Appendix

0.1 Formula for Gaussian Conditionals

A standard result of multivariate Gaussians states shows how to condition on a knowing a subset of the dimensions of a Gaussian vector. If

$$\begin{bmatrix} \mathbf{x}_A \\ \mathbf{x}_B \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_A \\ \boldsymbol{\mu}_B \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_A A & \boldsymbol{\Sigma}_A B \\ \boldsymbol{\Sigma}_B A & \boldsymbol{\Sigma}_B B \end{bmatrix} \right) \tag{1}$$

then

$$\mathbf{x}_A | \mathbf{x}_B \sim \mathcal{N}(\boldsymbol{\mu}_A + \boldsymbol{\Sigma}_{AB}\boldsymbol{\Sigma}_{BB}^{-1}\left(\mathbf{x}_B - \boldsymbol{\mu}_B\right), \boldsymbol{\Sigma}_{AA} - \boldsymbol{\Sigma}_{AB}\boldsymbol{\Sigma}_{BB}^{-1}\boldsymbol{\Sigma}_{BA}) \tag{2}$$

In the case of Gaussian processes, this result tells us how to condition on knowing the function values $[f(\mathbf{x}_1), f(\mathbf{x}_2), \ldots, f(\mathbf{x}_N)]$ at some subset of locations along the real line, indexed by $[\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]$.

0.2 Kernels

0.2.1 Base kernels

For scalar-valued inputs, the white noise (WN), constant (C), linear (Lin), squared exponential (SE), and periodic kernels (Per) are defined as follows:

$$\mathrm{C}(x,x') = \sigma^2 \tag{3}$$

$$\mathrm{WN}(x,x') = \sigma^2\delta(x-x') \tag{4}$$

$$\mathrm{Lin}(x,x') = \sigma^2(x-c)(x'-c) \tag{5}$$

$$\mathrm{SE}(x,x') = \sigma^2 \exp\left(-\frac{(x-x')^2}{2\ell^2}\right) \tag{6}$$

$$\mathrm{Per}(x,x') = \sigma^2 \frac{\exp\left(\frac{\cos\frac{2\pi(x-x')}{p}}{\ell^2}\right) - I_0\left(\frac{1}{\ell^2}\right)}{\exp\left(\frac{1}{\ell^2}\right) - I_0\left(\frac{1}{\ell^2}\right)} \tag{7}$$

$$\mathrm{CP}(k_1,k_2)(x,x') = \sigma(x)k_1(x,x')\sigma(x') + (1-\sigma(x))k_2(x,x')(1-\sigma(x')) \tag{8}$$

$$\boldsymbol{\sigma} = \sigma(x)\sigma(x') \tag{9}$$

$$\bar{\boldsymbol{\sigma}} = (1-\sigma(x))(1-\sigma(x')) \tag{10}$$

where $\delta_{x,x'}$ is the Kronecker delta function, I_0 is the modified Bessel function of the first kind of order zero and other symbols are parameters of the kernel functions.

0.2.2 Changepoints and changewindows

The changepoint, $\mathrm{CP}(\cdot,\cdot)$ operator is defined as follows:

$$\begin{aligned}\mathrm{CP}(k_1,k_2)(x,x') = \quad & \sigma(x)k_1(x,x')\sigma(x') \\ & +(1-\sigma(x))k_2(x,x')(1-\sigma(x'))\end{aligned} \tag{11}$$

where $\sigma(x) = 0.5 \times (1 + \tanh(\frac{\ell - x}{s}))$. This can also be written as

$$\mathrm{CP}(k_1,k_2) = \boldsymbol{\sigma}k_1 + \bar{\boldsymbol{\sigma}}k_2 \tag{12}$$

where $\boldsymbol{\sigma}(x,x') = \sigma(x)\sigma(x')$ and $\bar{\boldsymbol{\sigma}}(x,x') = (1-\sigma(x))(1-\sigma(x'))$.

Changewindow, $\mathrm{CW}(\cdot,\cdot)$, operators are defined similarly by replacing the sigmoid, $\sigma(x)$, with a product of two sigmoids.

0.2.3 Properties of the periodic kernel

A simple application of l'Hôpital's rule shows that

$$\mathrm{Per}(x, x') \to \sigma^2 \cos\left(\frac{2\pi(x - x')}{p}\right) \quad \text{as} \quad \ell \to \infty. \tag{13}$$

This limiting form is written as the cosine kernel (cos).

0.3 Model construction / search

0.3.1 Overview

The model construction phase of ABCD starts with the kernel equal to the noise kernel, WN. New kernel expressions are generated by applying search operators to the current kernel. When new base kernels are proposed by the search operators, their parameters are randomly initialised with several restarts. Parameters are then optimized by conjugate gradients to maximise the likelihood of the data conditioned on the kernel parameters. The kernels are then scored by the Bayesian information criterion and the top scoring kernel is selected as the new kernel. The search then proceeds by applying the search operators to the new kernel i.e. this is a greedy search algorithm.

In all experiments, 10 random restarts were used for parameter initialisation and the search was run to a depth of 10.

0.3.2 Search operators

ABCD is based on a search algorithm which used the following search operators

$$\mathcal{S} \quad \to \quad \mathcal{S} + \mathcal{B} \tag{14}$$

$$\mathcal{S} \quad \to \quad \mathcal{S} \times \mathcal{B} \tag{15}$$

$$\mathcal{B} \quad \to \quad \mathcal{B}' \tag{16}$$

where $\mathcal{S}$ represents any kernel subexpression and $\mathcal{B}$ is any base kernel within a kernel expression i.e. the search operators represent addition, multiplication and replacement.

To accommodate changepoint/window operators we introduce the following addi-

tional operators

$$\mathcal{S} \rightarrow \mathrm{CP}(\mathcal{S}, \mathcal{S}) \tag{17}$$
$$\mathcal{S} \rightarrow \mathrm{CW}(\mathcal{S}, \mathcal{S}) \tag{18}$$
$$\mathcal{S} \rightarrow \mathrm{CW}(\mathcal{S}, \mathrm{C}) \tag{19}$$
$$\mathcal{S} \rightarrow \mathrm{CW}(\mathrm{C}, \mathcal{S}) \tag{20}$$

where C is the constant kernel. The last two operators result in a kernel only applying outside or within a certain region.

Based on experience with typical paths followed by the search algorithm we introduced the following operators

$$\mathcal{S} \rightarrow \mathcal{S} \times (\mathcal{B} + \mathrm{C}) \tag{21}$$
$$\mathcal{S} \rightarrow \mathcal{B} \tag{22}$$
$$\mathcal{S} + \mathcal{S}' \rightarrow \mathcal{S} \tag{23}$$
$$\mathcal{S} \times \mathcal{S}' \rightarrow \mathcal{S} \tag{24}$$

where $\mathcal{S}'$ represents any other kernel expression. Their introduction is currently not rigorously justified.

0.4 Comparison of Predictive Accuracy

0.4.1 Interpolation

To test the ability of the methods to interpolate, we randomly divided each data set into equal amounts of training data and testing data. We trained each algorithm on the training half of the data, produced predictions for the remaining half and then computed the root mean squared error (RMSE). The values of the RMSEs are then standardised by dividing by the smallest RMSE for each data set i.e. the best performance on each data set will have a value of 1.

Figure 1 shows the standardised RMSEs for the different algorithms. The box plots show that all quartiles of the distribution of standardised RMSEs are lower for both versions of ABCD. The median for ABCD-accuracy is 1; it is the best performing algorithm on 7 datasets. The largest outliers of ABCD and spectral kernels are similar in value.

Changepoints performs slightly worse than MKL despite being strictly more gen-

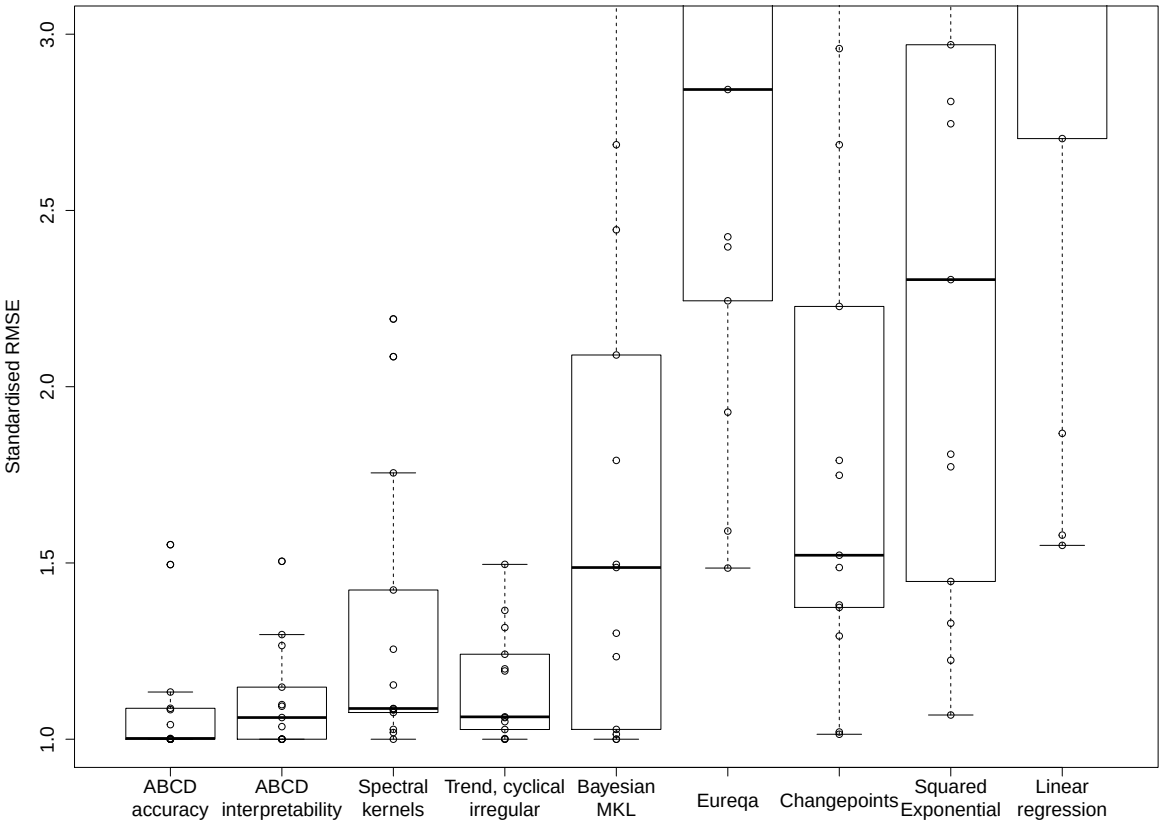

Fig. 1 Box plot of standardised RMSE (best performance = 1) on 13 interpolation tasks.

eral than Changepoints. The introduction of changepoints allows for more structured models, but it introduces parametric forms into the regression models (i.e. the sigmoids expressing the changepoints). This results in worse interpolations at the locations of the change points, suggesting that a more robust modeling language would require a more flexible class of changepoint shapes or improved inference (e.g. fully Bayesian inference over the location and shape of the changepoint).

Eureqa is not suited to this task and performs poorly. The models learned by Eureqa tend to capture only broad trends of the data since the fine details are not well explained by parametric forms.

0.4.2 Tabels of standardised RMSEs

See table 1 for raw interpolation results and table 2 for raw extrapolation results. The rows follow the order of the datasets in the rest of the supplementary material. The following abbreviations are used: ABCD-accuracy (ABCD-acc), ABCD-interpretability ((ABCD-int), Spectral kernels (SP), Trend-cyclical-irregular (TCI), Bayesian MKL (MKL), Eureqa (EL), Changepoints (CP), Squared exponential (SE) and Linear regression (Lin).

ABCD-acc	ABCD-int	SP	TCI	MKL	EL	CP	SE	Lin
1.04	1.00	2.09	1.32	3.20	5.30	3.25	4.87	5.01
1.00	1.27	1.09	1.50	1.50	3.22	1.75	2.75	3.26
1.00	1.00	1.09	1.00	2.69	26.20	2.69	7.93	10.74
1.09	1.04	1.00	1.00	1.00	1.59	1.37	1.33	1.55
1.00	1.06	1.08	1.06	1.01	1.49	1.01	1.07	1.58
1.50	1.00	2.19	1.37	2.09	7.88	2.23	6.19	7.36
1.55	1.50	1.02	1.00	1.00	2.40	1.52	1.22	6.28
1.00	1.30	1.26	1.24	1.49	2.43	1.49	2.30	3.20
1.00	1.09	1.08	1.06	1.30	2.84	1.29	2.81	3.79
1.08	1.00	1.15	1.19	1.23	42.56	1.38	1.45	2.70
1.13	1.00	1.42	1.05	2.44	3.29	2.96	2.97	3.40
1.00	1.15	1.76	1.20	1.79	1.93	1.79	1.81	1.87
1.00	1.10	1.03	1.03	1.03	2.24	1.02	1.77	9.97

Table 1 Interpolation standardised RMSEs

0.4.3 Comparison to Equation Learning

We now compare the descriptions generated by ABCD to parametric functions produced by an equation learning system. We show equations produced by Eureqa (**?**) for the data sets shown above, using the default mean absolute error performance metric.

The learned function for the solar irradiance data is

$$\text{Irradiance}(t) = 1361 + \alpha \sin(\beta + \gamma t) \sin(\delta + \epsilon t^2 - \zeta t)$$

where t is time and constants are replaced with symbols for brevity. This equation captures the constant offset of the data, and models the long-term trend with a product of sinusoids, but fails to capture the solar cycle or the Maunder minimum.

ABCD-acc	ABCD-int	SP	TCI	MKL	EL	CP	SE	Lin
1.14	2.10	1.00	1.44	4.73	3.24	4.80	32.21	4.94
1.00	1.26	1.21	1.03	1.00	2.64	1.03	1.61	1.07
1.40	1.00	1.32	1.29	1.74	2.54	1.74	1.85	3.19
1.07	1.18	3.00	3.00	3.00	1.31	1.00	3.03	1.02
1.00	1.00	1.03	1.00	1.35	1.28	1.35	2.72	1.51
1.00	2.03	3.38	2.14	4.09	6.26	4.17	4.13	4.93
2.98	1.00	11.04	1.80	1.80	493.30	3.54	22.63	28.76
3.10	1.88	1.00	2.31	3.13	1.41	3.13	8.46	4.31
1.00	2.05	1.61	1.52	2.90	2.73	3.14	2.85	2.64
1.00	1.45	1.43	1.80	1.61	1.97	2.25	1.08	3.52
2.16	2.03	3.57	2.23	1.71	2.23	1.66	1.89	1.00
1.06	1.00	1.54	1.56	1.85	1.93	1.84	1.66	1.96
3.03	4.00	3.63	3.12	3.16	1.00	5.83	5.35	4.25

Table 2 Extrapolation standardised RMSEs

The learned function for the airline passenger data is

$$\text{Passengers}(t) = \alpha t + \beta \cos(\gamma - \delta t)\text{logistic}(\epsilon t - \zeta) - \eta$$

which captures the approximately linear trend, and the periodic component with approximately linearly (logistic) increasing amplitude. However, the annual cycle is heavily approximated by a sinusoid and the model does not capture heteroscedasticity.