A. Appendix

A.1. Components Extracted from the Stock Market Data

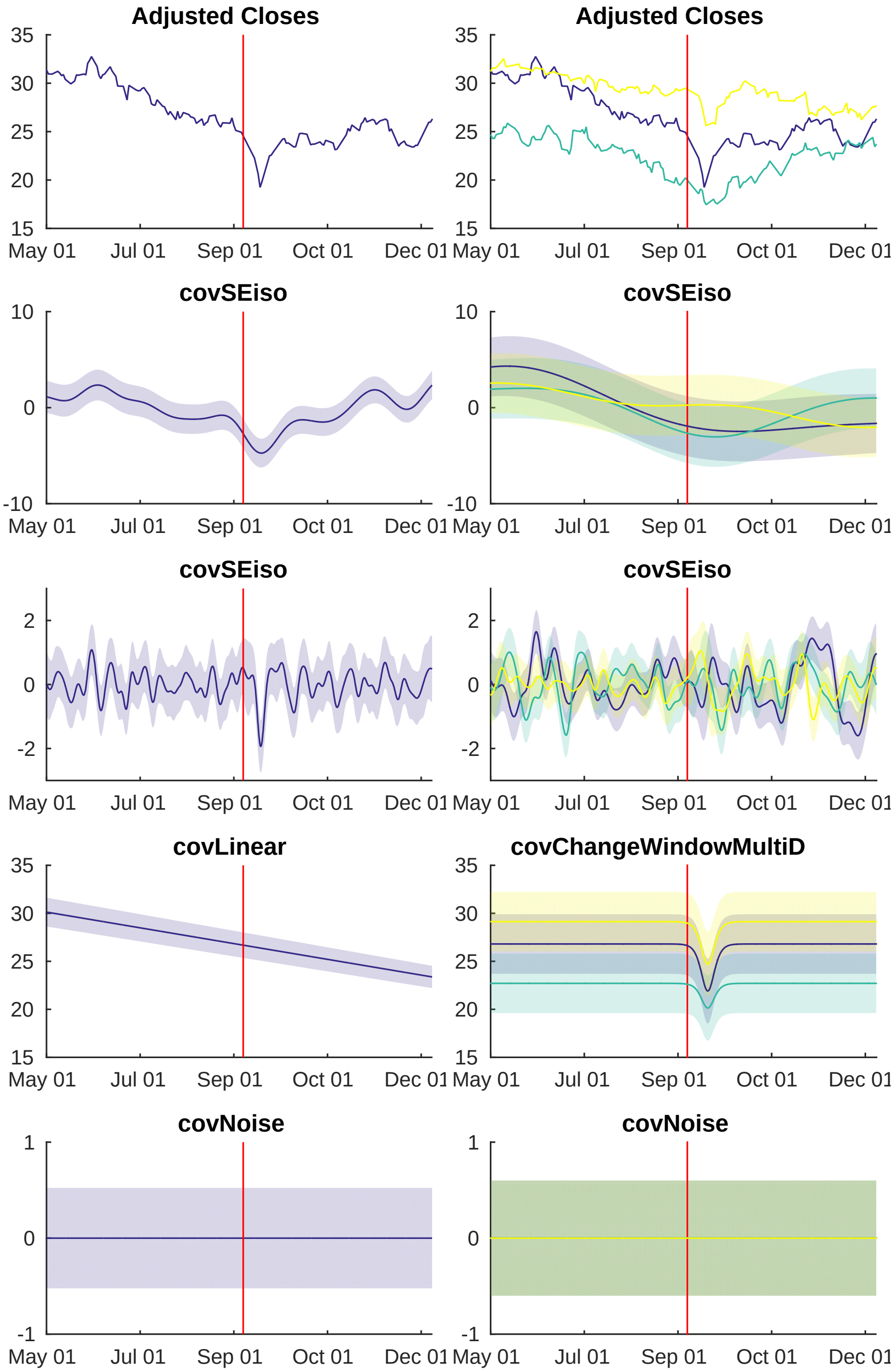

Figure 7: Figure that shows decomposition of adjusted closes of stock values. x axis is time and y axis is stock values. Red vertical line is at '2001/09/11' which is the day of 911. The third component of right side, 'covChangeWindowMultiD' captures sudden drop of stock values after the 911.

Time series	CKL	SRKL
GE	1.86	**0.48**
MSFT	3.60	**0.76**
XOM	1.18	**0.70**
PFE	1.63	**0.63**
C	1.10	**0.53**
WMT	1.84	**0.42**
INTC	1.22	**0.85**
BP	**1.01**	1.33
AIG	1.58	**0.94**

Table 2: Standardized RMSEs of stock data set

Time series	CKL	SRKL
New York	11.66	**9.97**
Los Angeles	2.75	**0.53**
Chicago	6.13	**1.24**
Pheonix	**3.39**	3.60
San Diego	7.94	**1.51**
San Francisco	2.65	**1.16**

Table 3: Standardized RMSEs of housing price data set

Time series	CKL	SRKL
IDR	3.55	**2.17**
MYR	**1.37**	2.35
ZAR	2.98	**1.32**
RUB	1.48	**1.29**

Table 4: Standardized RMSEs of currency exchange data set

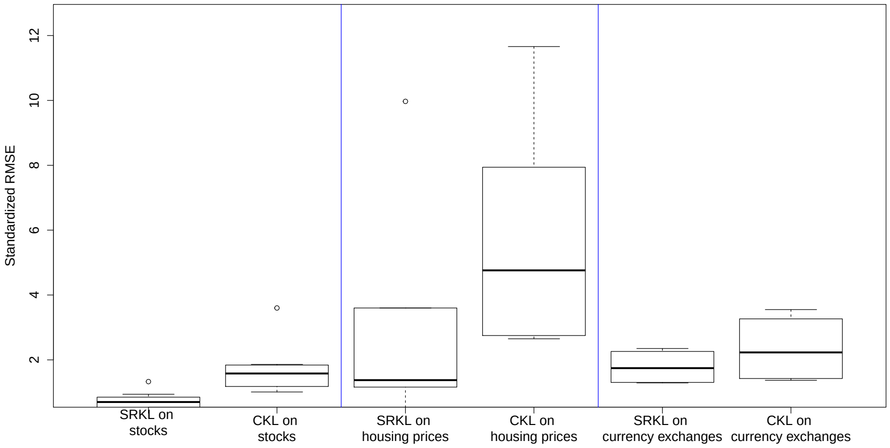

Figure 8: Box plot of standardized RMSEs on each data set

A.2. Details on standardized RMSEs

We provide the standardized RMSEs of all data sets in Table 2, Table 3, and Table 4. Moreover, Figure 8 shows the significant statistical improvement of SRKL in terms of extrapolation as discussed in the experiment section.

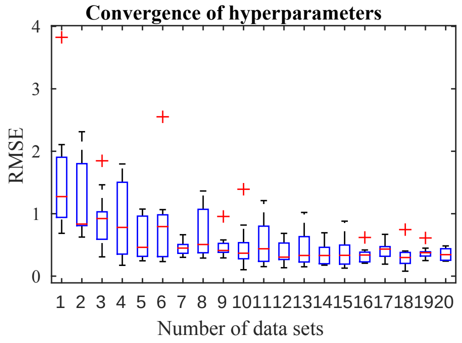

Figure 9: The convergence of error between the optimized hyperparameter and original true parameter that was used when generating data. The horizontal axis denotes the number of data sets used in hyperparameter optimization. The vertical axis shows the value of root mean squared error between optimized and true hyperparameters. Each boxplot shows the distribution of errors for each step.

Model-Stock		NLL	P	BIC
CKL	GE	116.36	7	266.75
	MSFT	111.17	6	251.51
	XOM	82.24	9	208.23
	PFE	62.91	12	184.14
	C	424.23	10	897.07
	WMT	149.96	5	324.23
	Total	**946.89**	49	2219.71
RKL-Total		1007.09	**18**	**2066.18**
SRKL-Total		988.15	204	3333.21

Table 5: The experimental comparisons of GP models with individual CKL, RKL and SRKL in the stock data set (top 6 US stocks in 2001). For each column, NLL means the negative log likelihood, P is the number of hyperparameters and BIC is the Bayesian information criterion. CKL is slightly better in a simple aggregation of NLL only. However, our RKL outperforms CKL in the BIC due to a significantly lower 21 parameters (P) compared to 49 parameters in CKL.

A.3. Learning Hyperparameters in Synthetic Data

First we had a covariance kernel function k that is already learned from the original ABCD algorithm. We generated 20 data sets from a GP prior with kernel k. Each data set has a different number of data points. After sampling the data, we tried to optimize the hyperparameters of k for $1, 2, \ldots, 20$ data sets chosen out of 20 sampled data sets, with the original hyperparameters as a starting point. We do this optimization 10 times for each number of data sets, from 1 to 20, with different combination of choice.

We calculated the root mean squared error (norm distance) between the true hyperparameter and the optimized hyperparameter vector. Figure 9 shows the result. The values along the horizontal axis means the number of data sets which are used in the optimization. Each boxplot shows the distribution of the error with 10 different results for each number of data sets used. It shows that as we increase the number of data sets in parameter optimization, the error between the optimized and the true hyperparameter decreases.

A.4. Experiment table with 6 stocks and 6 cities

Table 5 compares the fitness of the CKL, RKL and SRKL models. The results with line heading RKL are from our model. The others are from the original ABCD. The results with line heading Total is calculated by considering all of GP priors that are learned for each data set as a single model. So NLL and P values of individual models are added up and the total BIC of the CKL model as a single model was calculated using those values. In terms of negative log likelihood, applying ABCD individually gives better results. However if we compare the BIC, our model achieves better results, as our model has a reduced number of free parameters by sharing them through the data sets.

		RKL		SRKL		CKL	
Stocks	N	P	BIC	P	BIC	P	BIC
Top 3	387	**16**	**665.09**	108	1251.62	22	750.65
Top 6	774	**18**	**2066.18**	204	3333.21	49	2219.71
Top 9	1161	**32**	**3626.00**	300	5633.33	73	3985.03

Table 6: The BIC of CKL, RKL and SRKL in the stock data set. ‘Top 3’, ‘Top 6’ and ‘Top 9’ stocks were selected by their market capitalization ranks in 2011. As shown in Table 5, RKL requires fewer parameters than CKL. RKL models trained with 3 stocks and 6 stocks show better performance than individually optimized CKL models. When 9 stocks are considered, the individual CKL models show better performance than the single (shared) RKL model.

Model-City		NLL	P	BIC
CKL	New York	120.13	10	288.13
	Los Angeles	162.94	9	368.96
	Chicago	134.73	13	331.69
	Pheonix	101.76	11	256.17
	San Diego	155.64	12	368.73
	San Francisco	174.45	6	377.63
	Total	**849.64**	61	2100.62
RKL-Total		891.09	**33**	**1972.58**
SRKL-Total		1495.40	205	3707.94

Table 7: A comparison of BIC between individual CKL RKL and SRKL, with house data. For each column, NLL means negative log likelihood, P is the number of hyperparameters and BIC is the Bayesian information criterion. For each row, RKL is the result from our model. From New York to San Francisco, those results are from individual cities. Total is summation of those individual results.

Table 6 compares the fitness of the models with different numbers of data sets used in training. As the number of data sets increases, the number of parameters needed also increases. However RKL shows less need for parameters through the whole setup, since our model shares parameters though the data. RKL also performs well in terms of the BIC for TOP3 and TOP6. However as the number of data sets increases, the performance of the original CKL model surpasses our model, at TOP9. This is because our model fits the general structure but not individual specific ones. As the number of data sets increases, the learned structure cannot fully explain the individual specific patterns. And these accumulated errors in fitness, which is the NLL, offsets the advantages of the BIC which is from the reduced number of parameters.

Table 7 shows the fitness of models between our model and the original ABCD model. Similar to the stock market data case, the first line is our model and the others are from the original ABCD. Our model shows better results, a reduced number of free parameters and smaller BIC compared to Total case. However the original model still shows better fitness if we only consider NLL. From this experiment we can again confirm that the major contribution of the smaller BIC comes from the reduced number of parameters.

Table 8 compares results between our model and the ABCD for different numbers of data sets. Our model shows reduced numbers of parameters over all different sets of data since our model shares kernel parameters for multiple data sets. The

		RKL		SRKL		CKL	
SET	N	P	BIC	P	BIC	P	BIC
Top 2 cities	240	**11**	634.76	52	905.76	20	**634.00**
Top 4 cities	480	**18**	**1221.88**	134	3326.94	38	1424.18
Top 6 cities	720	**33**	**1876.47**	109	4339.54	61	2100.62

Table 8: The BIC of CKL, RKL and SRKL in the housing market data set. ‘Top 2’, ‘Top 4’ and ‘Top 6’ US cities were selected in terms of their city population rank. The BICs of the RKL models are similar or better than the BICs of individually trained CKL models.

original model shows better results in terms of BIC for the top 2 indices. However other than that, our model shows better results in terms of both number of parameters and BIC.