8th International Conference on Probabilistic Methods Applied to Power Systems, Iowa State University, Ames, Iowa, September 12-16, 2004

Gaussian Process Prior Models for Electrical Load Forecasting

Douglas J. Leith, Martin Heidl and John V. Ringwood *Senior Member, IEEE*

***Abstract*—This paper examines models based on Gaussian Process (GP) priors for electrical load forecasting. This methodology is seen to encompass a number of popular forecasting methods, such as Basic Structural Models (BSMs) and Seasonal Auto-Regressive Intergrated (SARI) as special cases. The GP forecasting models are shown to have some desirable properties and their performance is examined on weekly and yearly Irish load data.**

***Index Terms*—Electricity demand, Gaussian processes, load forecasting, modeling.**

I. INTRODUCTION

THE load forecasting problem [1] has attracted wide interest from a variety of research communities in an effort to provide increasingly accurate forecasts of demand. Accurate forecasts of demand are required, on a variety of timescales, in order to reduce spinning reserve, schedule maintenance and optimize energy trading mechanisms [2].

In the main, parametric structures are employed in load forecasting modes. For example, electrical load has been decomposed by many authors (e.g. [3], [4] and [5] to mention but a few) into a trend and cyclical component The presence of a trend and cyclical component in Irish load has also been examined in [6] and by Moutter *et. al*. [7] for a New Zealand utility. The above references utilize linear models, while a wide variety of nonlinear models have also been employed, including the addition of nonlinear autoregressive components [8], Hammerstein/Wiener models [9] and a wide range of techniques based on artificial neural networks as exemplified in [10], [11] and [12]. Some load forecasting techniques have also been reported which use a statistical/inference approach, rather than that of a fixed parametric model structure, e.g. [13] and [14].

For time series models, where the underlying parametric structure is largely unknown, one of the significant challenges is to determine an appropriate form of parameterisation for the forecasting model. Some implementations, such as neural network models, are claimed to be 'non-parametric', since there is a generalized set of basis functions, which are combined in a linear, or nonlinear, way. However, the particular form of the basis functions must be chosen, as must the finite set of weights and biases, which strengthen the parametric typing of the system. One could muse, however, that if an infinite set of basis functions and parameters were used, then the structure would be truly non-parametric.

One class of model, which can reasonably be classed as non-parametric, is a model based on Gaussian process (GP) priors, which can be considered (in some sense) to have equivalence to a model based on an infinite set of nonlinear basis functions [15]. The central idea here is, without parameterising the model, to place a *prior* directly on the space of functions. This can be thought of as the generalization of a Gaussian distribution over a finite vector space to a function space of infinite dimension. Then, rather then specifying the parameters of the model, the GP is specified by its mean and covariance function, where the covariance function has a particular (but simple) parametric structure, enumerated by a set of *hyperparameters*. While GPs have been used in time series forecasting [16], to the best of these authors' knowledge, this is the first application of GP to the electrical load forecasting problem. There are a number of features that make GPs potentially attractive in a load forecasting context, including:

- Models can be determined using a relatively small number of points (unlike neural networks), which makes them potentially useful in forecasting of annual load,
- Co-variance functions are modular – they can easily be synthesized from components representing particular features in the data e.g. seasonality,
- They extend in a very seamless way from linear to nonlinear synthesis,
- Though not specifically recommended for extrapolation, they perform is a safe manner when asked to extrapolate outside the training data, and
- Confidence intervals are easy to evaluate, which help to indicate where model is unreliable (e.g. lack of training data).

However, GPs are not without their difficulties and, in common with neural networks, the determination of the (hyper) parameters requires the solution of a non-convex optimization problem. Also, in spite of the fact that GPs are described by relatively few hyperparameters, their dimensionality is determined by the amount of training data, which possibly presents some problems, since the covariance matrix must be inverted in order to perform a prediction. These issues are dealt with in the paper in the context of the development of GP-based forecasting models for weekly and annual Irish electrical load.

II. Gaussian Process Models

Consider a stochastic process with output $y \in \Re$ conditional on input $\mathbf{z} \in \Re^q$. Suppose we have N measurements of input-output pairs, $\{(\mathbf{z}_i, y_i)\}_{i=1}^N$, and denote these by M. We are interested in using this data to learn the posterior probability distribution of y at some arbitrary input value $\mathbf{z}$; that is, $p(y|\mathbf{z}, M)$. The following exposition of GPs is necessarily brief, with the reader directed to [18] for a more complete treatment.

A Covariance Formulation

Consider initially conventional Bayesian parametric modelling approaches to solving the special case where $y(\mathbf{z})=f(\mathbf{z})+v$ *i.e.* a smooth scalar function $f(\mathbf{z})$ with additive Gaussian white measurement noise v, that is, the classical regression task. For example, say we believe that $f(\mathbf{z})$ has the form $\Psi(\mathbf{z})\boldsymbol{\theta}$ with:

$$\Psi(\mathbf{z}) = \begin{bmatrix} \varphi_1(\mathbf{z}) & \cdots & \varphi_p(\mathbf{z}) \end{bmatrix} \tag{1}$$

and $\boldsymbol{\theta}$ the model parameters (that is, $f(\mathbf{z})$ consists of a weighted combination of fixed basis functions $\varphi_i(\mathbf{z}), i=1..p$). We have that:

$$p(y|\mathbf{z}, M) = \int p(y|\mathbf{z}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|M) d\boldsymbol{\theta} \tag{2}$$

where $p(\boldsymbol{\theta}|M)$ is the probability distribution over the set of possible models. Bayes Rule states that:

$$p(\boldsymbol{\theta}|M) = p(M|\boldsymbol{\theta}) p(\boldsymbol{\theta}) / p(M) \tag{3}$$

where the likelihood $p(M|\boldsymbol{\theta})$ embodies the information provided by the measured data, the prior $p(\boldsymbol{\theta})$ embodies our prior beliefs regarding the process and p(M) is simply a normalising factor which is hereafter ignored. Assuming a Gaussian prior, then we have:

$$p(\boldsymbol{\theta}|M) \propto \underbrace{\exp\left[-\frac{1}{2}\left(\mathbf{Y}-\hat{\mathbf{Y}}\right)^T \boldsymbol{\Lambda}_v^{-1}\left(\mathbf{Y}-\hat{\mathbf{Y}}\right)\right]}_{p(M|\boldsymbol{\theta})} \underbrace{\exp\left[-\frac{1}{2}\left(\boldsymbol{\theta}-\bar{\boldsymbol{\theta}}\right)^T \boldsymbol{\Lambda}_\theta^{-1}\left(\boldsymbol{\theta}-\bar{\boldsymbol{\theta}}\right)\right]}_{p(\boldsymbol{\theta})} \tag{4}$$

where:

$\mathbf{Y} = [y_1 \ldots y_N]^T$ is the vector of measured outputs

$\hat{\mathbf{Y}} = \begin{bmatrix} \Psi(\mathbf{z}_1)\boldsymbol{\theta} & \cdots & \Psi(\mathbf{z}_N)\boldsymbol{\theta} \end{bmatrix}^T$

$\boldsymbol{\Lambda}_v$ is the measurement noise covariance

$\bar{\boldsymbol{\theta}}$ and $\boldsymbol{\Lambda}_\theta$ the mean and covariance of the prior resp.

with large variance specified when we have little confidence in our prior knowledge. Commonly, it is assumed that the probability distribution $p(y|\mathbf{z},M)$ is sharply peaked so that:

$$p(y|\mathbf{z}, M) \approx p(y|\mathbf{z}, \boldsymbol{\theta}_{MAP}) \propto \exp\left[-\frac{1}{2}(y-\Psi(\mathbf{z})\boldsymbol{\theta}_{MAP})^T \boldsymbol{\Lambda}_v^{-1}(y-\Psi(\mathbf{z})\boldsymbol{\theta}_{MAP})\right] \tag{5}$$

where $\boldsymbol{\theta}_{MAP}$ is the value of the parameter vector $\boldsymbol{\theta}$ for which $p(\boldsymbol{\theta}|M)$ is maximal. The mean prediction from this model (*i.e.* the fit to the function $f(\mathbf{z})$) is therefore $\Psi(\mathbf{z})\boldsymbol{\theta}_{MAP}$, with variance $\boldsymbol{\Lambda}_v$. Notice that owing to linearity dependence of the output on the parameters and because the prior distribution of the parameters $\boldsymbol{\theta}$ is Gaussian, the prior probability distribution of the observations $p(y(\mathbf{z}_1), y(\mathbf{z}_2), \ldots, y(\mathbf{z}_n))$ is also Gaussian for any set of inputs $\{\mathbf{z}_k \in \Re^q, k=1,..n\}$ and any n. For example, when the prior on the parameters is mean zero, it is readily verified that $p(y(\mathbf{z}_1), y(\mathbf{z}_2))$ is mean zero with covariance, $\mathrm{cov}(y(\mathbf{z}_i), y(\mathbf{z}_j))$:

$$C(y(\mathbf{z}_i), y(\mathbf{z}_j)) = \begin{bmatrix} \Psi(\mathbf{z}_1)\boldsymbol{\Lambda}_\theta\Psi^T(\mathbf{z}_1) & \Psi(\mathbf{z}_1)\boldsymbol{\Lambda}_\theta\Psi^T(\mathbf{z}_2) \\ \Psi(\mathbf{z}_2)\boldsymbol{\Lambda}_\theta\Psi^T(\mathbf{z}_1) & \Psi(\mathbf{z}_2)\boldsymbol{\Lambda}_\theta\Psi^T(\mathbf{z}_2) \end{bmatrix} \tag{6}$$

A multivariate normal joint prior $p(y(\mathbf{z}_1), y(\mathbf{z}_2), \ldots, y(\mathbf{z}_n))$ is the defining property of a Gaussian Process prior (GP) model. The multi-variate normal distribution is characterised by its mean (assumed zero in the sequel for convenience although this may be readily relaxed) and covariance. Evidently, the foregoing basis function model is one example of a GP model, corresponding to a specific choice of covariance function. In general, we use a GP model to carry out inference as follows. We have that:

$$p(y|\mathbf{z}, M) = p(y, \{y_i\}_{i=1}^N | \mathbf{z}, \{\mathbf{z}_i\}_{i=1}^N) / p(M) \tag{7}$$

where $p(\{y_i\}_{i=1}^N, y | \{\mathbf{z}_i\}_{i=1}^N, \mathbf{z})$ is the prior and p(M) acts simply as a normalising constant and so can be ignored here. Hence, substituting for our Gaussian prior

$$p(y|\mathbf{z}, M) \propto \exp\left[-\frac{1}{2}\begin{bmatrix} y & \mathbf{Y} \end{bmatrix}^T \begin{bmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{12}^T & \Lambda_{22} \end{bmatrix}^{-1} \begin{bmatrix} y & \mathbf{Y} \end{bmatrix}\right] \tag{8}$$

where $\Lambda_{11}=C(y(\mathbf{z}),y(\mathbf{z}))$, the ij^{th} element of the covariance matrix $\boldsymbol{\Lambda}_{22}$ is equal to $C(y(\mathbf{z}_i),y(\mathbf{z}_j))$ and the i^{th} element of vector Λ_{12} equals $C(y(\mathbf{z}),y(\mathbf{z}_i))$. Applying the partitioned matrix inversion lemma [17], it follows that

$$p(y|\mathbf{z}, M) \propto \exp\left[-\frac{1}{2}(y-\hat{y}(\mathbf{z}))^T \Lambda^{-1}(\mathbf{z})(y-\hat{y}(\mathbf{z}))\right] \tag{9}$$

where

$$\hat{y}(\mathbf{z}) = \boldsymbol{\Lambda}_{12}^T \boldsymbol{\Lambda}_{11}^{-1} \mathbf{Y}, \ \Lambda(\mathbf{z}) = \boldsymbol{\Lambda}_{22} - \boldsymbol{\Lambda}_{12}^T \boldsymbol{\Lambda}_{11}^{-1} \boldsymbol{\Lambda}_{12} \tag{10}$$

The mean prediction from this model is therefore $\hat{y}(\mathbf{z})$, with variance $\Lambda(\mathbf{x})$. Note that $\hat{y}(\mathbf{z})$ is simply a weighted linear combination of the measured data points, $\mathbf{Y}$, using weights $\boldsymbol{\Lambda}_{12}^T \boldsymbol{\Lambda}_{11}^{-1}$.

B Popular Examples of GP Prior Models

The Gaussian process prior formulation used here is not new [18], but does provide a useful unifying framework which encompasses a broad spectrum of popular models. Within a regression context special cases of GP models include, for example, linear regression models, basis function models (including those with infinitely many basis functions), splines and certain types of multi-layer perceptron [15,16,19,20]. It is also clear that a GP is a Bayesian form of kernel regression model [21,22]. The GP framework also encompasses the case of linear dynamics with Gaussian process and observation noise (and thereby Kalman filters/smoothers). This is of particular relevance to the forecasting context and may be seen

as follows[1]. Consider the standard linear stochastic state-space equations:

$$\begin{aligned} \mathbf{x}_{i+1} &= \mathbf{F}_i \mathbf{x}_i + \mathbf{G}_i \mathbf{u}_i \\ \mathbf{y}_i &= \mathbf{H}_i \mathbf{x}_i + \mathbf{v}_i \end{aligned} \tag{11}$$

with $\mathbf{x}_o$, $\mathbf{u}_i$, $\mathbf{v}_i$ are uncorrelated zero mean Gaussian processes such that:

$$\text{cov}(\mathbf{u}_i,\mathbf{u}_j)=\mathbf{Q}_i\delta_{ij},\ \text{cov}(\mathbf{v}_i,\mathbf{v}_j)=\mathbf{R}_i\delta_{ij},\ \text{cov}(\mathbf{x}_o,\mathbf{x}_o)=\mathbf{\Pi}_o \tag{12}$$

This stochastic state-space formulation is the basis for a number of popular approaches used in forecasting, backcasting, smoothing and signal extraction including the Basic Structural Model (BSM) approach of Harvey [23], the Dynamic Harmonic Regression method of Young et al. [24]. It follows from the linearity of the dynamics (11) that the state $\mathbf{x}$ and output $\mathbf{y}$ are normally distributed with zero mean and covariance:

$$\mathbf{\Pi}_{i+1} = \mathbf{F}_i \mathbf{\Pi}_i \mathbf{F}_i^T + \mathbf{G}_i \mathbf{Q}_i \mathbf{G}_i^T \tag{13}$$

where $\mathbf{\Pi}_i=E(\mathbf{x}_i\mathbf{x}_i^T)$. Hence,

$$E(\mathbf{x}_i \mathbf{x}_j^T) = \begin{cases} \mathbf{\Phi}(i,j)\mathbf{\Pi}_j & i \geq j \\ \mathbf{\Pi}_i^T \mathbf{\Phi}(j,i)^T & i < j \end{cases} \tag{14}$$

and

$$E(\mathbf{y}_i \mathbf{y}_j^T) = \begin{cases} \mathbf{H}_i \mathbf{\Phi}(i,j) \mathbf{\Pi}_j \mathbf{H}_j^T & i > j \\ \mathbf{H}_i \mathbf{\Pi}_i \mathbf{H}_i^T + \mathbf{R}_i & i = j \\ \mathbf{H}_i \mathbf{\Pi}_i^T \mathbf{\Phi}(j,i)^T \mathbf{H}_j^T & i < j \end{cases} \tag{15}$$

where

$$\mathbf{\Phi}(i,j) = \begin{cases} \mathbf{F}_{i-1} \cdots \mathbf{F}_j & i > j \\ \mathbf{I} & i = j \end{cases} \tag{16}$$

is the state transition matrix. Evidently, the linear state space model is simply a Gaussian process prior model with covariance function defined by (13). It is readily verified using the partitioned matrix inversion lemma [17] that the various forms of recursive Kalman smoother equations (with Kalman filter forward pass followed by fixed interval smoothing backward pass) for calculating the posterior mean and covariance are precisely equivalent to the en bloc formulation (10). Of course, the special structure of the covariance function (16) is exploited in the Kalman smoother formulation to reduce the computational burden associated with the calculations. Specifically, while inversion of the covariance matrix in (10) generally requires $O(N^3)$ operations, where N is the number of measured data values, covariance matrices associated with the covariance function (16) can be inverted in $O(N^2)$ operations; see, for example, [25].

[1] An equivalence between Kalman smoothers associated with integrated random walk models and a certain en bloc kernel regression approach has, for example, been previously noted by Young & Pedregal (1999).

III. Load Forecasting – Weekly Data

Ireland operates an island network with demand coming from a mixture of domestic, commercial and industrial users, with a peak load of 3800 MW. It is strongly driven by weather inputs, particularly temperature, at the weekly level. A number of studies have examined the load forecasting problem for this data and can provide a benchmark for the current study, especially since a number of them fall into the sub-categories of GPs indicated in Section II B.

A. Irish Load Data

A total of 679 weekly data points are available, shown in Fig. 1.

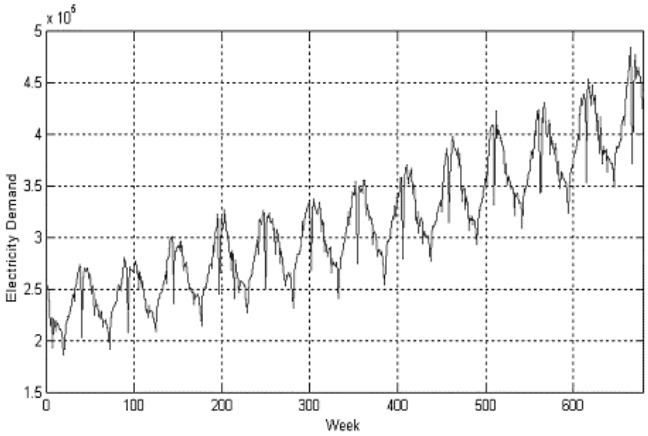

Fig.1 Weekly Irish Load Data

The strong seasonality of the data is evident, along with special annual events, such as Christmas and Summer vacation.

B Covariance Function Synthesis

Since every sum of valid covariance functions is a valid covariance function, it is possible to define elemental, or primitive, covariance functions that can be combined as required. (Covariance functions can also be generated from various products of simpler covariance functions). Elemental covariance functions relevant to the present forecasting context include the following.

1) Smooth Nonlinearity/Trend

$$C(y(\mathbf{x}_i), y(\mathbf{x}_j)) = a \ \exp[-\frac{1}{2} \sum_{k=1}^{D} \frac{(x_i^k - x_j^k)^\alpha}{r_k}] \tag{17}$$

where x_i^k denotes the k^{th} element of vector $\mathbf{x}_i$. This covariance function simply expresses that the outputs associated with input values close to each other have a higher correlation than outputs associated with input values far away from each other. There is length scale parameter r_k associated with each input which characterises the distance in that direction over which observations are correlated. The parameter $\alpha \in (0,2]$ determines the rate of decay, while parameter a defines the vertical scale of variations of a typical function.

2) Linear Component

$$C(y(\mathbf{x}_i), y(\mathbf{x}_j)) = \sum_{k=1}^{D} w_k \; x_i^k \; x_j^k \tag{18}$$

It is straightforward to verify that this covariance function is equivalent to the linear model:

$$y(\mathbf{x}_i) = \sum_{k=1}^{D} m_k \; x_i^k \tag{19}$$

where the prior distributions, $p(m_k)$, k=1..D, are Gaussian zero mean with covariance w_k and it is assumed that w_i, w_j are independent for $i \neq j$.

3) White Noise

$$C(y(\mathbf{x}_i), y(\mathbf{x}_j)) = v \; \delta(i, j) \tag{20}$$

where

$$\delta(i, j) = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \tag{21}$$

4) Periodic/Seasonal Component

$$C(y(\mathbf{x}_i), y(\mathbf{x}_j)) = c \; \exp\left[-\frac{1}{2} \sum_{k=1}^{D} \left(\frac{\sin(\frac{\pi}{\lambda_k} (x_i^k - x_j^k))}{r_k} \right)^2 \right] \tag{22}$$

This models a function that is periodic with period λ_k in the k^{th} input direction. Similarly to the nonlinear/trend component described above, r_k determines the length scale over which observations are correlated within a single period. See Fig.2 for an illustration of this covariance function.

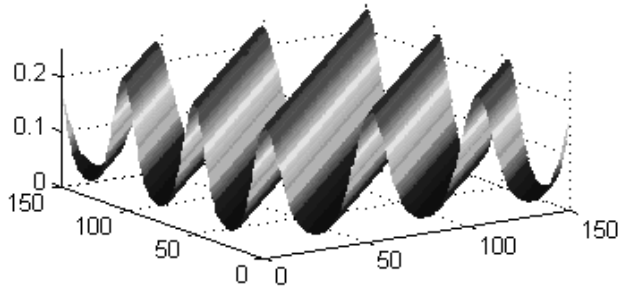

Fig.2: Covariance matrix representing seasonal data

With regard to the Irish weekly load forecasting context, the presence of a trend, a seasonal component and a white noise component is evident by inspection (it remains of course to establish if other elements might be present). The following covariance function is therefore postulated:

$$C(y(\mathbf{x}_i), y(\mathbf{x}_j)) = \underbrace{a \; \exp\left[-\frac{1}{2} \frac{(x_i^1 - x_j^1)^2}{r_1} \right]}_{\text{trend}} + \underbrace{c \exp\left[-\frac{1}{2} \left(\frac{\sin(\frac{\pi}{\lambda} (x_i^1 - x_j^1))}{r} \right)^2 \right]}_{\text{seasonal component}}$$
$$+ \underbrace{a \; \exp\left[-\frac{1}{2} \frac{(x_i^2 - x_j^2)^2}{r_2} \right]}_{\text{autocorrelation}} + \underbrace{v \delta(i, j)}_{\text{noise}} \tag{23}$$

where the vector of regressors, $\mathbf{x}_i = [i \quad y_{i-1}]^T$. The regressor vector and hyperparameters of the covariance are selected to maximise the likelihood of the training data (a simple gradient descent optimisation). As an example, the weekly forecast for a full year is shown Fig.3. It can be seen that predictable features, such as the Christmas week, are captured very accurately by the GP model. Enumerated forecast results for this GP model are presented in Table 1.

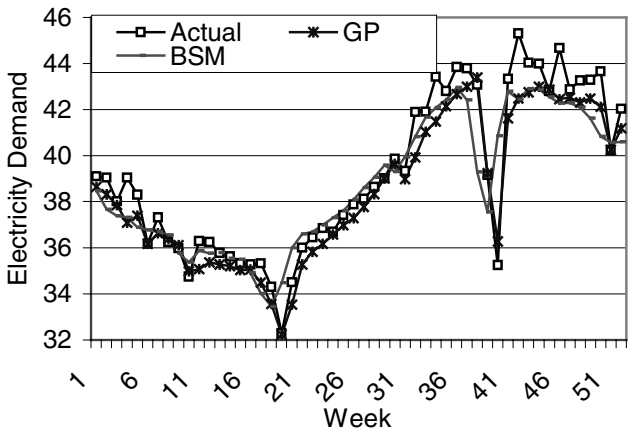

Fig.3: One Year ahead predictions

Table 1: GP results compared with BSM and SARI models

Year	SARI		BSM		GP	
	MAPE (%)	MSPE (%)	MAPE (%)	MSPE (%)	MAPE (%)	MSPE (%)
1996	2.69	3.22	3.89	4.52	2.18	2.87
1997	7.57	8.14	4.93	5.63	1.91	2.52
1998	6.65	6.97	2.42	2.99	1.84	2.46
1999	1.81	2.32	2.57	3.69	1.84	2.34
2000	2.03	2.77	3.08	4.27	2.12	2.56
Ave.	**4.15**	**4.68**	**3.38**	**4.22**	**1.98**	**2.55**

Forecasting accuracy is assessed by making one year ahead predictions (52 weeks) over five consecutive years i.e. for a dataset containing N years of data, the first N-5 years are used as training data to forecast the N-4th year, then the first N-4 years are used as training data to forecast the N-3rd year and so on. The metrics used are the Mean Absolute Percentage Error (MAPE) and the Mean Square Percentage Error (MSPE):

$$\text{MAPE} = \sum_{i=1}^{N} \frac{\left| e_i / a_i \cdot 100 \right|}{N}, \quad \text{MSPE} = 100 \cdot \sqrt{\sum_{i=1}^{N} \frac{\left(e_i / a_i \right)^2}{N}} \tag{24}$$

It can be seen that the GP model consistently generates significantly more accurate forecasts than either the SARI (Seasonal AutoRegressive Integrated) or BSM models developed in previous studies [26,27] for Irish weekly load data.

IV. Load Forecasting – Yearly Data

A. Irish Load Data

Fig.4 shows the load expressed as total electricity sales (TS) over 41 years, plotted with Irish Gross Domestic Product (GDP) as a candidate explanatory variable. In general, the data can be characterized as having a rising trend with a sharp change in slope towards the latter years.

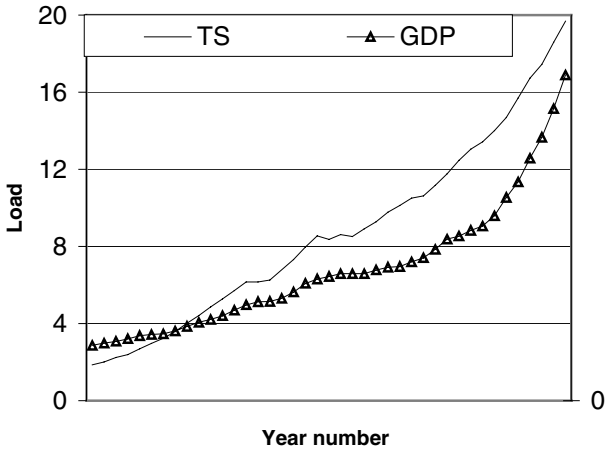

Fig.4: Irish annual load data characteristics

Initially, the use of a linear regression model for forecasting annual electricity sales in Ireland is studied (such models are in widespread use for such data). Running five year forecasts are considered i.e. a forecast is made for the 5 years following a section of training data. As before, the mean and the standard deviation of the MAPE and the MSPE are employed as metrics. The forecast period encompasses the period of major change in the growth rate of the Irish economy. Consequently, it is hoped that a method capable of providing accurate forecasts for this period may be able to assist in forecasting sales in future years where continuing economic change is anticipated (forecasts of GDP corresponding to a variety of economic scenarios are available from the Economic and Social Research Institute Ireland). Table 2 shows the mean and standard deviation of MAPE and MSPE for various choices of regressor, where y_i represents the load in year i.

Table 2: MAPE and MSPE for different regressor choices

Regressor		[i]	[i GDP]	[i GDP y_{i-1}]	[i GDP y_{i-1} y_{i-2}]
MAPE (%)	mean	7.88	4.54	4.87	6.68
	std	3.93	3.05	2.90	5.00
MSPE (%)	mean	8.61	5.20	6.20	8.37
	std	3.92	3.64	3.42	6.05

However, a more revealing result is the variation in the prediction error with the forecast period, shown in Fig.5. Evidently, the forecasts rapidly become inaccurate following the economic changes mentioned.

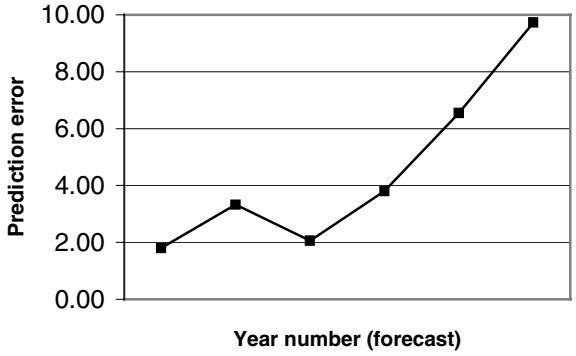

Fig.5: Change in prediction error with date of forecast

Some insight into this behaviour can be gained by noting that while the rate of growth of GDP increases abruptly at a certain point, a corresponding change in electricity sales is not apparent. This is further verified by partitioning the data into two periods and noting that linear regression models individually fitted to the data in these two periods have substantially different parameter values. It appears, therefore, that the annual electricity data has a non-stationary/heteroscedastic character. Forecasting in such cases is well known to be challenging.

B. Covariance Function Synthesis

The standard form of linear regression model is also, of course, an example of a GP model (see Section II B). The requirement is for a more appropriate choice of covariance function which better reflects the character of the data. The sales data is primarily a trend, with no periodic component. It seems, moreover, that the trend can be approximately partitioned into two regions each with linear dependence on time and GDP, although a single linear model is not valid over the entire dataset. That is, we require a linear model with varying slope and offset of the form:

$$y(\mathbf{x}_i) = \sum_{k=1}^{D} m_k(\mathbf{x}_i)\, x_i^k + b_k(\mathbf{x}_i) + \eta_i \qquad (25)$$

where the slope m_k and offset b_k vary with the explanatory variable $\mathbf{x}_i$ and η_i is Gaussian white noise. Assuming that the slope and offset parameters are independent with zero mean Gaussian prior distributions we have that:

$$C\big(y(\mathbf{x}_i), y(\mathbf{x}_j)\big) = \sum_{k=1}^{D} C\big(m_k(\mathbf{x}_i), m_k(\mathbf{x}_j)\big)\, x_i^k x_j^k + C\big(b_k(\mathbf{x}_i), b_k(\mathbf{x}_j)\big) + v\,\delta(i,j) \qquad (26)$$

where v is the noise covariance. The variation of the slope and offset could be modelled in many ways, but for simplicity we initially use the previous nonlinear/trend elemental covariance and assume that the slope and offset vary with time but not with GDP. That is, with regressor vector, $\mathbf{x}_i = [i \quad GDP_i]^T$:

$$C(y(\mathbf{x}_i), y(\mathbf{x}_j)) = \exp[-\frac{1}{2}\frac{(x_i^1 - x_j^1)^2}{r_m}] \sum_{k=1}^{D} w_k\ x_i^k\ x_j^k + a\ \exp[-\frac{1}{2}\frac{(x_i^1 - x_j^1)^2}{r_b}] + v\ \delta(i,j) \quad (27)$$

In the limiting case where r_m, $r_b \to \infty$ a standard linear regression model (see) is recovered. The hyperparameters of the covariance are selected to maximise the likelihood of the training data (a simple gradient descent optimisation). The fit for the training time gets much better with this new covariance function. The MAPE during the training time, if all data is used, is 2.67% with the linear model and 1.63% with the new model.

V. Conclusion

Gaussian processes have been shown to be not only a viable forecasting methodology but also a potential unifying framework within which many existing forecasting methodologies may be cast. In many cases, existing methods such as structural models, SARI models and, indeed, neural networks, can be looked on as special cases of the general GP model. As such, in their most general form, GPs are likely to produce superior results, providing a suitable form of covariance function can be found. In the forecasting examples considered, the form of covariance function can be readily determined from a cursory inspection of the data. However, two potential problems arise with the use of GPs:

- The determination of the hyperparameters of the covariance function represents, in general, a non-convex optimization problem, and
- The use of the GP model requires the determination of a matrix inverse at each forecasting step, as evidenced in (10), with the dimension of the inverse dependent on the number of training points used.

For the applications considered, the above were not found to be problematic, but arguably the requirement for a significant matrix inverse in the case of the annual data obviates the need for a (relatively) complex GP model implementing a reasonably simple linear model. One other difficulty, in the particular case of the annual load, is the length of the forecast duration relative to the amount of training data available. Typically, forecasts of 5 to 15 years are required, demanding that the GP model operate well outside the space of the training data, with resulting poor performance, since the GP model forecast decays to zero as one moves away from the training data. Parametric models do not exhibit this feature, though extrapolation outside the training data carries no guarantees of model validity. Some comment is also worthy in relation to the comparison of a nonlinear GP model with the linear BSM and SARI for the weekly data. The use of neural network-based SARI and BSM models improved the MAPE performance by approx 1% [27], which would still not wipe out the advantage of GPs demonstrated in Table 1.

References

[1] E.D. Bunn and E.D. Farmer (Eds.) *Comparative Models for Electrical Load Forecasting*, Wiley, New York, 1985.

[2] A.P. Douglas, A.M. Breipohl, F.N. Lee and R. "Risk due to load forecast uncertainty in short term power planning", *IEEE Trans. On Power Systems*, Vol.13, No.4, 1998, pp 1493-1499.

[3] S.L. Chen and , F.C Kao, "An efficient algorithm to model and forecast hourly weather sensitive load", *Journal of the Chinese Institute of Electrical Engineering*, Vol.3, No.3, 1996, pp 231-243.

[4] R. Ramanathan, R. Engle, C.W.J Granger, , F.V. Araghi and C. Brace, "Short-run forecasts of electricity loads and peaks", *International Journal of Forecasting*, Vol.13, 1997, pp 161-174.

[5] T. Haida, and S. Muto, "Regression based peak load forecasting using a transformation technique", *IEEE Transactions on Power Systems,* Vol.9, No.4, 1994, pp 1788-1794.

[6] F. Murray and J.V. Ringwood, "Improvement of electricity consumption forecasts using temperature inputs", *Simulation Practice and Theory*, Vol.2, No.3, 1994, pp 121-139.

[7] S.P. Moutter, B.E. Bodger, P.T. Gough, "Spectral decomposition and extrapolation of variations in electricity loading", *IEE Proceedings-Part C*, Vol.113, 1986, pp 247-255.

[8] H. Mori and Kobayashi, " Optimal fuzzy inference for short-term load forecasting", *IEEE Transactions on Power Systems,* Vol.11, No.1, 1996, pp 390-396.

[9] Q.C. Lu, W.M. Grady, M.M. Crawford, and G.M. Anderson, "An adaptive nonlinear predictor with orthogonal escalator structure for short-term load forecasting", *IEEE Transactions on Power Systems,* Vol.4, No.1, 1989, pp 158-164.

[10] G.A. Darbellay and M. Slama, " Forecasting the short-term demand for electricity, do neural networks stand a better chance?", *International Journal of Forecasting*, Vol.16, 2000, pp 71-83.

[11] M.M. Elkateb, K. Solaiman, K. and Y.Al-Turki, "A comparative study of medium-weather-dependant load forecasting using enhanced artificial/fuzzy neural network and statistical techniques", *Neurocomputing*, Vol.23, 1998, pp 3-13.

[12] J.V. Ringwood, F.T. Murray and D. Bofelli, "Forecasting electricity demand on short, medium and long time scales using neural networks", *Journal of Intelligent and Robotic Systems* , Vol. 31, Nos. 1-3, May-July 2001.

[13] W. Charytoniuk,, M.S. Chen, P. Kotas and P. Van Olinda, "Demand forecasting in power distribution systems using non-parametric probability density estimation", *IEEE Transactions on Power Systems,* Vol.14, No.4, 1999, pp 1200-1206.

[14] S. Kiartzis, A. Kehagias, A. Bakirtzis and V. Petridis, "Short term load forecasting using a Bayesian combination method", *Electrical Power and Energy Systems*, 19 (3), 1997, pp 171-177.

[15] D. MacKay, "Gaussian processes: a replacement for supervised neural networks ?", in *Advances in Neural Information Processing Systems*, MIT Press, 1997.

[16] C. K. I. Williams, "Prediction with Gaussian Processes: From Linear Regression to Linear Prediction and Beyond", In "Learning and Inference in Graphical Models", ed. M. I. Jordan, Kluwer, 1998.

[17] R. Horn and C. Johnson, *Matrix Analysis*, Cambr. Univ. Press, 1996.

[18] A. O'Hagan, "On curve fitting and optimal design for regression", *J. Royal Stat. Soc. B*, Vol.40, 1978, pp 1-42.

[19] R. Neal, *Bayesian Learning for Neural Networks*, Springer, 1996.

[20] M. Gibbs, *Bayesian Gaussian Processes for Regression and Classification*, PhD Thesis, Cambridge University, 1997.

[21] P.J. Green and B.W. Silverman, *Nonparametric Regression and Generalised Linear Models*, Chapman and Hall, 1994.

[22] N. Christianini, N and J. Shawe-Taylor, *An Introduction to Support Vector Machines*, Cambridge University Press, 2000.

[23] A.C. Harvey, *Forecasting, Structural Time Series Models and the Kalman Filter*, CRC Press, 1989.

[24] P.C. Young, D. Pedregal and W. Tych, "Dynamic harmonic regression", *Journal of Forecasting*, Vol.18, 1999, pp 369-394.

[25] T. Kailath and A.H. Sayed, *Fast Reliable Algorithms for Matrices with Structure*, SIAM, 1999.

[26] F. Murray and , J.V. Ringwood, "Improvement of electricity consumption forecasts using temperature inputs", *Simulation Practice and Theory*, Vol.2, No.3, 1994, pp 121-139.

[27] J.V. Ringwood and F. Murray, "Forecasting of weekly electricity consumption using neural networks", *Proc. Irish DSP and Control Colloquium (IDSPCC '96)*, Dublin, June 1996.