Discussion on Regression and Classification using Gaussian process priors

Pietro Muliere
Università degli Studi di Pavia, Italy.

1 General comments

Regression is one of the most common data problems and numerous methods exist for tackling it. A nonparametric Bayesian regression model must be based on a prior distribution over the infinite-dimensional space of possible regression functions. In this paper Neal examines the Gaussian process approach to regression and classification.

Assume we have observed data for n cases: $(\mathbf{x_1}, y_1), \ldots, (\mathbf{x_n}, y_n)$ where $\mathbf{x_i} = (x_{i1}, \cdots, x_{ik})$ is the vector of k inputs, for $i = 1, \ldots, n$. We wish to construct a model for this data and then to make predictions using the model. For example, given the vector $\mathbf{x_{n+1}}$ we may want to predict y_{n+1}.

The generative model for our data is, for every n,

$$y_n = f(\mathbf{x_n}) + v_n$$

where $f(\mathbf{x_n})$ is our modeling function and v_n is some noise. In order to specify the prior process the Author uses a Gaussian process since *a Gaussian process is completely described by its mean value vector and its covariance function.*

The Bayesian model thus defined is very similar to others that have appeared in the literature (see: Kimeldorf and Wahba (1970a,1970b), Wahba (1978), Wecker and Ansley (1983), Cox (1993)). Related models and applications are discussed in Diaconis (1988).

An analogous approach is followed by Liptser and Shyryayev (1972) for discussing the Kalman-Bucy filter in the discrete and continuos case.

It is obvious that the conceptual simplicity of Gaussian processes make them very attractive in many applied problems. In my opinion the role of Gaussian processes used as priors in regression or classification problems is well understood from the methodological point of view. Of course this does not imply that they are always computationally easy to use; this seems particularly true when the covariance function depends on unknown parameters. With respect to this side of the problem the paper by Neal looks like a valuable contribution for the researcher interested in applications.

In the rest of my discussion I will try to put Neal's work in perspective by focusing on a few theoretical questions related to the use of Gaussian processes in Bayesian inference.

2 How to select the prior

In practical applications the covariance function selected may or may not be an adeguate description of our state of knowledge. Consequently I argue that a Gaussian process seems inappropriate when the form of the covariance function is hard to justify on the basis of what is known about the observables. Roughly speaking it should be the problem that suggests the form of the prior.

Along this principle, a different approach in selecting the prior considers de Finetti style theorems which characterize models in terms of invariance. The idea is that the statistician begins the model building phase by postulating reasonable symmetries for the distribution of observables. The theorems then give simple descriptions of all the distributions with the given symmetries. Diaconis, Eaton and Lauritzen (1992) developed such theorems leading to invariance justifications of the usual linear model, analysis of variance, and some covariance models in multivariate analysis.

In some other practical applications the form of the predictive distribution may be an adeguate description of our state of knowledge. Therefore, the prior is characterized via the sequence of predictive distributions. For some examples of this approach see Regazzini(1998) in this conference and the references therein.

3 Covariance functions and regression models

A wide variety of covariance functions can be used in the Gaussian process framework. Obviously different covariance functions will determine different regression models. In order to elicit a covariance function one wants to know what sort of regression model it generates.

The problem thus becomes: *Is it possible to determine the regression model starting from the covariance function ?*

I will give two answers to this question.

3.1 Canonical expansions of random functions

Any random function $Y(t)$ can be expressed in the form:

$$Y(t) = m(t) + \sum_{v} V_v x_v(t) \tag{1}$$

where: $m(t)$ is the mathemathical expectation of the random function $Y(t)$, V_v are uncorrelated random variables with zero expected values, and the functions $x_v(t)$ are some non random functions. Representation (1) is called canonical expansion of $Y(t)$; the variables V_v are the coefficients of the canonical expansion while $x_v(t)$ are coordinate functions. In the general case the canonical expansion of a random function is an infinite series but in particular cases it can be a finite sum. Expressing the random function $Y(t)$ by the canonical expansion (1), the covariance function of the variables $Y(t)$ and $Y(t')$ is given by:

$$Cov(Y(t), Y(t')) = \sum_{v} D_v x_v(t) \bar{x}_v(t'). \tag{2}$$

where D_v are the dispersions of the random functions V_v.

Any representation of a covariance function by (2) is called a canonical expansion of the covariance function. Thus the canonical expansion (2) of the covariance function of a random variable $Y(t)$ follows from the canonical expansion (1). *Conversely it is possible to show that the canonical expansion (1) of a random function follows from the canonical expansion (2) of the covariance function.* For the construction of a canonical expansion for a random function from the canonical expansion of its covariance function see, for example, Chapter 9 of Pugachev (1965).

3.2 Markovian representation of stochastic processes

Given a sequence of covariance matrices $\{R_l, l = 0, 1, \ldots\}$ of a zero-mean d-dimensional stationary stochastic process $\{y_n\}$ one can find a Markovian representation of y_n which is given by

$$\begin{aligned} x_{n+1} &= Fx_n + w_n \\ y_n &= Hx_n \end{aligned}$$

where the dimension of x_n,w_n,F and H are $k \times 1, k \times 1, k \times k$, and $p \times k$, respectively, and w_n is an k-dimensional zero mean white noise. Akaike (1975) has shown that if the covariance function $R_l = Cov(y_n, y_{n+l}), l = 0, 1, \ldots$, admits a finite dimensional factorization

$$R_l = HF^lG$$

then the process y_n has two specific Markovian representation. The state x_n of one of these representations is defined as a set of mutually orthogonal random variables which contains the full information of the past of the process to be expressed by the present and the future, and the state of the other representation contains the full information of the future of the process to be expressed by the present and past. This fact is proved by using the concept of canonical variables which is well developed in the field of multivariate statistical analysis.

The problem of finding all minimal Markovian (state space) representations of a given random process is known as the *stochastic realization problem.* In this problem one is given a stochastic process and asked to construct a stochastic system in a specified class such that the output of this system equals the given process. The practical motivation of this problem comes from communications and control, econometrics, time series analysis, and other areas where model building is important. The stochastic realization problem for Gaussian processes has been extensively investigated by Lindquist,Picci and Ruckebusch (1979) and reference therein. The minimal splitting subspace can be interpreted as a subspace of smallest size containing all the information from the past needed in predicting the future and all the information in the future required to estimate the past. This fact make it an obvious candidate for a state space.

4 Covariance function and spectral density

Bayesian inference of a time series generated by a stationary, discrete-time, Gaussian stochastic process is possible parametrizing the spectral density.

Let Y(t) be a discrete time, stationary stochastic process. Given a covariance function, R, it can be shown that there exixts a unique, real-valued, non-decreasing, right continuous function F defined on $[-\pi, \pi]$ such that

$$Cov(Y(t), Y(t')) = R(j) = \int_{-\pi}^{\pi} exp(iwj)dF(w)$$

where $j = 0, \pm 1, \ldots$, $F(-\pi) = 0$ and F has increments symmetric about zero. F is called the spectral distribution function. If the spectral distribution function is absolutely continuous with respect to Lebesgue measure on $[-\pi, \pi]$ then given the covariance function, R, one can find an almost-everywhere unique, real -valued, symmetric (about 0), non-negative function f defined on $[-\pi, \pi]$ such that

$$R(j) = \int_{-\pi}^{\pi} exp(iwj)f(w)dw$$

where $j = 0, \pm 1, \ldots$. The function f is the spectral density function of R or of the process. The converse of the above statement can be shown to be true also.

One common way to proceed with the analysis of a time series is to concentrate on the analysis of the spectral density of the generating process. A Bayesian appraoch based on the spectral density for stationary Gaussian processes is analyzed by Shore (1980).

5 The Stochastic model

In Neal's paper the inputs $\mathbf{x_1}, \mathbf{x_2}, \cdots, \mathbf{x_n}$ are fixed and it is not clear how to elicit the covariance function when they are a sample of random vector $\mathbf{X} = (X_1, \ldots, X_k)$. In fact, in this case, we need to make assumptions on the joint distribution of $(\mathbf{X}, y)$.

I would like to suggest a simple nonparametric approach for solving the problem of prediction by means of the most used process in Bayesian nonparametrics :the Dirichlet process.

Consider a sample of n observations $(\mathbf{x_1}, y_1), \ldots, (\mathbf{x_n}, y_n)$ of the random variables $(\mathbf{X}, Y)$ with joint probability distribution F. Let the form of F be uncertain. Consider then a future observation$((X_{n+1,1}, \cdots, X_{n+1,k}, Y_{n+1}))$, the $n+1$-th , from F and suppose we wish to make inference about Y_{n+1}, modeling the relationship between the set of the explanatory variables and the response variables Y_{n+1} given the past observed data. Cifarelli, Muliere and Scarsini (1981), following an idea of Goldstein (1976), defined a good linear predictor assuming F is a Dirichlet process prior. Therefore they search the Bayes predictor within the class of decision rules of the form

$$\sum_{i=1}^{k} \beta_i X_{n+1,i}$$

which minimizes

$$E[(Y_{n+1} - \sum_{i=1}^{k} \beta_i X_{i,n+1})^2 | \text{data}],$$

where the expectation is taken with respect to the predictive distribution.

We indicate with D the matrix of order $k \times k$ with elements

$$d_{ij} = E[X_{n+1,i} X_{n+1,j} | \text{data}].$$

and with $\mathbf{b}$ the $k \times 1$ vector with elements

$$b_i = E[X_{r,n+1} Y_{n+1} | \text{data}].$$

Then the optimal linear predictor is given by

$$\beta^* = D^{-1}\mathbf{b}.$$

Assume now that the prior for the random quantity F is a Dirichlet process: that is $F \in \mathcal{D}(MF_0)$ where $M > 0$ is a real number and $F_0 = F_0(x_1, x_2, \cdots, x_k, y)$ is a proper distribution function. F_0 can be interpreted as a prior guess at F whereas M is the 'measure of faith' in this guess. We thus obtain:

$$d_{ij} = \frac{M}{M+n} E_{F_0}(X_i X_j) + \frac{n}{M+n} (\frac{1}{n} \sum_{t=1}^{n} x_{ti} x_{tj})$$

where E_{F_0} is the expected value with respect to the initial guess F_0 for F. Hence, denoting with $p = \frac{M}{M+n}$, the matrix D assumes the following form:

$$D = pE_{F_0}(\mathbf{X}'\mathbf{X}) + (1-p)\frac{\mathbf{x}'\mathbf{x}}{n}$$

where $\mathbf{x}$ is the $n \times k$ matrix whose rows are the vectors $\mathbf{x_i}, i = 1, \ldots, n$.

In the same way we derive the vector $\mathbf{b}$:

$$\mathbf{b} = pE_{F_0}(\mathbf{X}Y) + (1-p)\frac{\mathbf{x}'\mathbf{y}}{n}$$

with $\mathbf{y}' = (y_1, \ldots, y_n)$. Note that if $M \to \infty$ we obtain $\beta^* = (E_{F_0}(\mathbf{X}'\mathbf{X}))^{-1}E_{F_0}(\mathbf{X}Y)$ whereas if $M \to 0$ we obtain $\beta^* = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{y}$, the ordinary least square estimator.

For a generalization of this approach see Luscia (1983) and Poli (1985).

6 Computation and Implementation

I would also like to comment on the practical implementation of the proposed approach, particularly with reference to the computational aspects. I will make a few short remarks.

- As I have already pointed out, it is important to understand which type of prior beliefs on the regression function correspond to a specific covariance function. Within the proposed approach, this becomes quite difficult, due to the necessity of several integrations, both over the hyperparameters and the latent variables. I thus suggest the Author to investigate, at least empirically, the input-target relationships embodied in the proposed models (such as the three-way classification example), for instance by means of MCMC sampling from the prior over the hyperparameter space.

- In practical applications, there will be a potentially large amount of explanatory inputs. It seems that the proposed model has computational difficulties to deal with this situation, essentially due to the calculation of a large number of derivatives of the loglikelihood function as well as to the need to tune empirically many stepsizes in the "leapfrog updates". These problems would persist, although in a different form, even using a random walk Metropolis-Hastings approach. Thus, it seems likely that some form of pre-processing of the data, perhaps using frequentist-based model selection methods is needed, inconsistently with the claimed full-fledged bayesianity of the approach. Alternatively, one may wonder whether the proposed approach could be simplified, considering, for instance, a Markovian representation of the stochastic process, possibly allowing a higher amount of "local" computations.

- The proposed prior on the hyperparameters assumes their independence or, at best, their exchangeability according to an unspecified grouping label. I wonder whether one should consider richer priors, for instance, including in the covariance function a dependence on input interactions. This would compensate for (often realistic) collinearity effects, as well as allow some forms of "borrowing strength" when doing quantitative learning on the hyperparameters. Moreover, from a computational viewpoint, a richer, possibly modular, prior would possibly allow many computations to localise, thus improving computational efficiency. I wonder whether the architecture of a neural network would be of help in this respect, as it is for graphical models (see e.g. Lauritzen, 1996).

- It would be desirable to develop some more formal methods of model comparison and criticism. Is this possible under the current approach ? Would not it be better to consider, instead of the full parameter space, a varying-dimension parameter space, according to which variablles are relevant under the current model ? I wonder if, using the latter approach, in conjunction with an appropriate Monte Carlo simulation method, such as the reversible jump MCMC (Green, 1995) the computational efficiency of the Monte Carlo simulations can be improved.

7 Final remarks

We conclude reminding that inference and predictions problems in linear models, analysis of variance, discriminant analysis are solved using a generalization of a Dirichlet process called " *mixture of products of Dirichlet* *processes*" (see: Cifarelli (1979), Cifarelli, Muliere and Scarsini (1981), Consonni (1981), Muliere and Scarsini (1983)).

References

Akaike,H. (1975), Markovian representation of stochastic processes by canonical variables.*SIAM,J. Control*,13,162-173.

Cifarelli,D.M. (1979), Impostazione bayesiana di un problema di analisi della varianza con approccio non parametrico. *Quaderni Istituto di Matematica Finanziaria* Università di Torino.

Cifarelli,D.M. Muliere,P. and Scarsini,M.(1981), Il modello lineare nell'approccio Bayesiano non parametrico. *Quaderni dell'Istituto Matematico "G. Castelnuovo"* Università degli Studi,Roma.

Consonni, G. (1981), Impostazione Bayesiana di un problema di analisi discriminatoria nell'ambito di un modello non parametrico. *Rivista di matematica per le Scienze Economiche e Sociali* 4, 89-102.

Cox,D.D. (1993). An analysis of Bayesian inference for nonparametric regression.*Ann.Statist.* 21,903-923.

Diaconis, P. (1988). Bayesian numerical analysis. In *Statistical Decision Theory and Related Topics 4* (S.G. Gupta and J.O. Berger,eds.) 1, 173-175, Springer,Berlin.

Diaconis, P. Eaton,M.L. and Lauritzen,S.L. (1992), Finite de Finetti Theorems in Linear models and Multivariate Analysis. *Scand. J. Statist.* 19, 289-315.

Goldstein,M. (1976). Bayesian analysis of regression problems. *Biometrika* 63, 51-58.

Green, P.J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination, *Biometrika*, 82, 711–732.

Kimeldorf, G.and Wahba,G. (1970a). A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *Ann.Math.Statist.* 41, 495-502.

Kimeldorf, G.and Wahba,G. (1970b). Spline functions and stochastic processes. *Sankhya,Ser.A* 32, 173-180.

Lauritzen, S.L. (1996). *Graphical models.* Oxford University Press, Oxford.

Lindquist,A. Picci G. and Ruckebusch,G. (1979). On minimal splitting subspaces and markovian representations. *Math. Systems Theory,* 12, 271-279.

Liptser, R.S. and Shyryayev A.N. (1972). Statsitics of conditionally Gaussian random sequences. *Proc.Sixth. Berkeley Symposium Math. Statistics and Probablity (1970)*,vol. II, Univeristy of California press, 389-422.

Luscia,F. (1983), Impiego di tecniche bayesiane non parametriche nel modello lineare generale. *Giornale degli Economisti e Annali di Economia* 79-89.

Muliere,P and Scarsini M. (1983). Impostazione Bayesiana di un problema di analisi della varaianza a due criteri. *Giornale degli Economisti e Annali di Economia,* 519- 526.

Poli, I. (1985), A Bayesian nonparametric estimate for multivariate regression. *Journal of Econometrics*28, 171-182.

Pugachev, V.S. (1965). *Theory of random Functions and its Application to Control problems.* Pergamon Press.

Regazzini,E. (1998). Old and recent results on the relationship between predictive inference and statistical modelling either in nonparametric or parametric form. In this conference.

Shore,R.W. (1980), A Bayesian approach to the spectral analysis of stationary time series. In *Bayesian Analysis in Economic Theory and Time series analysis* (C.A. Holt,Jr. and R.W. Shore, eds.) North-Holland.

Wahba,G.(1983). Bayesian confidence intervals for the cross-validated smoothing spline.*J.Roy.Statist.Soc.Ser.B* 45,133-150.

Wecker,W.and Ansley,C. (1983). The signal extraction approach to nonlinear regression and spline smoothing.*J. Amer.Statist.Assoc.* 78, 81-89.