Bayesian Regression Mixtures of Experts for Geo-Referenced Data

Gerhard Paaß and Jörg Kindermann

Fraunhofer Institute for Autonomous Intelligent Systems (AIS)
53754 St. Augustin, Germany
{Paass, Kindermann}@ais.fraunhofer.de

Abstract. Politicians, planners and social scientists have an increasing need for tools clarifying the spatial distribution of relevant features. Special interest is in what-if analyses: what would happen if we change some features in a specific way. To predict future developments requires a statistical model with inherent modelling uncertainty. In this paper we investigate Bayesian models which on the one hand are able to represent complex relations between geo-referenced variables and on the other hand estimate the inherent uncertainty in predictions. For solution the models require Markov-Chain Monte Carlo techniques.

1 Introduction

Spatial interpolation and extrapolation is an essential feature of many Geographic Information Systems (GIS). It is a procedure for estimating values of a variable at un-sampled locations. Based on Tobler's Law of Geography, which stipulates that observations close together in space are more likely to be similar than those farther apart, these procedures try to separate spatial correlation from random noise. They can, however, be divergent and lead to very different results if the underlying structural assumptions are not fulfilled. As a consequence, an understanding of the initial assumptions and methods used is key to the spatial interpolation process.

Bayesian statistics offers a way to mitigate these problem. It describes the uncertainties inherent in a statistical analysis by means of probability distributions, which capture the degree of belief that a quantity is located in some interval. This applies to observable quantities like the variables of interest as well as to unobservable quantities as the parameters of models, and their structural properties. During the last decade a number of new computation strategies have been developed which allow the solution of large scale problems for very complex models by means of stochastic simulation.

In this paper we describe the Bayesian variant of a flexible semi-parametric model, a *mixture of experts*, which is able to represent a wide variety of complex dependencies. It is composed of a series of localized component models called *experts,* which cover local properties of the relation in question.

In the next chapter we will describe spatial data and their specific properties. In chapter three we shortly describe classical statistical inference procedures like

least squares and in chapter four its Bayesian counterparts. Chapter five compiles some ensemble methods which use collections of possible models to describe the inherent variability or to get better predictions by forming a committee. Chapter six describes the classical methods of spatial statistics, which mostly are derived from linear least squares approaches. The last chapter is central to the paper as it analyses different advanced nonlinear procedures and assesses their potential in the spatial domain, especially in a Bayesian framework.

2 Bayesian Statistics

2.1 Basic Setup

Bayesian inference is the process of fitting a probability model to a set of data and summarizing the result by a probability distribution on the parameters of the model. In addition probability distributions on unobserved quantities such as predictions for new observations may be derived. Assume we have independent observations $(z_1, \mathbf{x}_1), \ldots, (z_n, \mathbf{x}_n)$ of the inputs $\mathbf{x}_i \in \Re^k$ and outputs $z_i \in \Re$. We may arrange the observed inputs in the matrix $\mathbf{X} = \mathbf{X}_{(n,k)} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)'$ and the outputs in a vector $\mathbf{z} = \mathbf{z}_{(n,1)} = (z_1, \ldots, z_n)'$. Bayesian inference assumes the existence of a joint distribution $p(\theta, \mathbf{z}, \mathbf{X})$. We are especially interested in the conditional distribution $p(\theta, \mathbf{z}|\mathbf{X}) = p(\theta|\mathbf{X})p(\mathbf{z}|\theta, \mathbf{X})$. Let the *prior distribution* $p(\theta) = p(\theta|\mathbf{X})$ describe the information about the parameter θ before the data $\mathbf{z}$ is available.

Then *Bayes' rule* yields the *posterior density*

$$p(\theta|\mathbf{z}, \mathbf{X}) = \frac{p(\theta, \mathbf{z}|\mathbf{X})}{p(\mathbf{z}|\mathbf{X})} = \frac{p(\theta|\mathbf{X})p(\mathbf{z}|\theta, \mathbf{X})}{p(\mathbf{z}|\mathbf{X})} = \frac{p(\theta)p(\mathbf{z}|\theta, \mathbf{X})}{\int p(\theta)p(\mathbf{z}|\theta, \mathbf{X})\, d\theta} \tag{1}$$

which describes the distribution of parameters after $\mathbf{X}$ and $\mathbf{z}$ have been observed. To make predictive inferences about an unknown observable and a new input $\mathbf{x}_0$ we calculate the prediction $p(z|\mathbf{x}_0, \theta)$ for each θ and $\mathbf{x}_0$ and weight them according to the posterior $p(\theta|\mathbf{z}, \mathbf{X})$ of parameters

$$p(z|\mathbf{x}_0, \mathbf{z}, \mathbf{X}) = \int p(z|\mathbf{x}_0, \theta)p(\theta|\mathbf{z}, \mathbf{X})d\theta \tag{2}$$

This gives us the complete distribution of z for a new input $\mathbf{x}_0$ in the light of the data $\mathbf{z}, \mathbf{X}$. We can evaluate any characteristics of this distribution, for instance its expected value $E(z|\mathbf{x}_0, \mathbf{z}, \mathbf{X})$, or a *highest posterior density region* which is the smallest region covering the output with a prescribed probability, e.g. 90%. It may – of course – no longer be contiguous but consist of a set of contiguous subset.

2.2 Prediction and Markov Chain Monte Carlo

The predictive distribution for the output of interest conditional on the new input $\mathbf{x}_0$ and the observed data $\mathbf{z}, \mathbf{X}$ was $p(z|\mathbf{x}_0, \mathbf{z}, \mathbf{X}) = \int p(z|\mathbf{x}_0, \theta)p(\theta|\mathbf{z}, \mathbf{X})d\theta$.

We may approximate the integral by a sum

$$p(z|\mathbf{x}_0, \mathbf{z}, \mathbf{X}) \approx \frac{1}{N} \sum_{j=1}^{N} p(z|\mathbf{x}_0, \theta_j) \qquad \theta_j \sim p(\theta|\mathbf{z}, \mathbf{X}) \tag{3}$$

If the θ_j are independently generated according to the posterior then the sum converges to the desired density by the law of large numbers. Subsequently we may describe $p(z|\mathbf{x}_0, \mathbf{z}, \mathbf{X})$ by different features, e.g. expectation, variance or posterior intervals.

The *Metropolis-Hastings algorithm* allows to generate a sample of parameter values θ_j distributed according to the posterior density. This involves the construction of a Markov chain $\theta(0), \theta(1), \ldots$ designed to be distributed according to the posterior density $p(\theta|\mathbf{z}, \mathbf{X})$. If the chain is currently at $\theta = \theta(t)$, the *Metropolis-Hastings algorithm* [Tie94] requires a *proposal density* $\boldsymbol{q}(\theta, \tilde{\theta})$, which is the conditional distribution of proposing a move from θ to $\tilde{\theta}$. The *acceptance probability* is defined as

$$p_{\mathrm{acc}}(\theta, \tilde{\theta}) = \min \left\{ 1, \frac{p(\tilde{\theta}|\mathbf{z}, \mathbf{X})\, \boldsymbol{q}(\tilde{\theta}, \theta)}{p(\theta|\mathbf{z}, \mathbf{X})\, \boldsymbol{q}(\theta, \tilde{\theta})} \right\} \tag{4}$$

With probability $p_{\mathrm{acc}}(\theta, \tilde{\theta})$ the candidate $\tilde{\theta}$ is accepted and the chain moves to $\theta(t+1) = \tilde{\theta}$. Otherwise the candidate is rejected and $\theta(t+1)$ takes the old value θ. For the actual transition probability $\boldsymbol{p}(\theta, \tilde{\theta}) := \boldsymbol{q}(\theta, \tilde{\theta}) p_{\mathrm{acc}}(\theta, \tilde{\theta})$ the *detailed balance* condition holds for all $\theta, \tilde{\theta}$

$$p(\theta|\mathbf{z}, \mathbf{X})\, \boldsymbol{p}(\theta, \tilde{\theta}) = p(\tilde{\theta}|\mathbf{z}, \mathbf{X})\, \boldsymbol{p}(\tilde{\theta}, \theta) \tag{5}$$

If the resulting Markov chain is aperiodic and irreducible (i.e. reaches all states with positive probability) then its distribution converges to an invariant stationary limit distribution, which is just the posterior distribution $p(\theta|\mathbf{z}, \mathbf{X})$ [Tie94].

If we have several candidate models, where the number and the interpretation of parameters is different, the approach cannot be used. [Gre95] has proposed an MCMC-scheme for varying dimension problems, termed *reversible jump MCMC*. When the current state is θ and $p(\theta|\mathbf{zX})$ is the target probability measure (the posterior density) we consider a countable number of different moves m. Depending on the state θ a move m and a destination $\tilde{\theta}$ is proposed with $q_m(\theta, \tilde{\theta})$ as joint distribution. $q_m(\theta, \tilde{\theta})$ may be a sub-probability measure, with probability $1 - \sum_m \int_{\tilde{\theta}} q_m(\theta, \tilde{\theta}) d\tilde{\theta}$ no move is attempted.

For the case that θ and $\tilde{\theta}$ have the same dimension, the procedure reduces to the Metropolis-Hastings algorithm (4). Now suppose that starting from θ a move of type m is proposed that yields a higher-dimensional $\tilde{\theta}$. This can be implemented by drawing a vector $\mathbf{u}$ of continuous variables distributed according to a known density $p_m(\mathbf{u})$ independent of θ. It is required that the sum of the dimensions of θ and $\mathbf{u}$ is equal to the dimension of $\tilde{\theta}$. Then the new state $\tilde{\theta}$ is defined by an invertible deterministic function $\tilde{\theta} = h_m(\theta, \mathbf{u})$. The reverse of

the move can be accomplished by using the inverse transformation, so that the proposal is deterministic. Then we get the acceptance probability

$$p_{\mathrm{acc}m}(\theta, \tilde{\theta}) = \min\left(1, \left|\frac{\partial h_m(\mathbf{u}, \theta)}{\partial(\mathbf{u}, \theta)}\right| * \frac{p(\tilde{\theta}|\mathbf{z}, \mathbf{X})\, j_m(\tilde{\theta})}{p(\theta|\mathbf{z}, \mathbf{X})\, j_m(\theta)\, p_m(\mathbf{u})}\right) \tag{6}$$

Here $j_m(\theta)$ and $j_m(\tilde{\theta})$ are the probabilities of selecting move m or its inverse in states θ and $\tilde{\theta}$ respectively. [Gre95] shows that the detailed balance condition 5 holds and consequently the equilibrium distribution of the resulting Markov chain is the posterior distribution $p(\theta|\mathbf{z}, \mathbf{X})$. Similar to the usual Metropolis-Hastings formula 4 the densities have to be known only up to a factor, which cancels out in 6.

The reversible jump algorithm is a major improvement in the Markov Chain Monte Carlo approach. It allows to explore complete model classes instead of a single model with a given structure. Note, however, that for the different classes prior probabilities are required.

Instead of specifying all priors explicitly we may use mixtures between priors of different shapes, so called hierarchical models [GCSR95, p.119], to introduce the prior information in a less restrictive way. The final weighting of different priors then is determined by the data.

3 Mixtures of Experts

Modular and hierarchical systems allow complex learning problems to be solved by dividing the problem into a set of subproblems, each of which may be simpler to solve than the original problem. In spatial statistics it is natural to assume that the data can be well described by a collection of functions, each of which is defined over a relatively local region of the input space. A *modular architecture* can model such data by allocating different modules to different regions of the space. *Hierarchical architectures* arise when we assume that the data are well described by a multi-resolution model – a model in which regions are recursively divided into subregions. An example is the decision tree model.

The learning algorithm simultaneously has to determine a partition of the input space into regions as well as the local models (experts) within each region. The *mixture of experts* approach developed by [JJNH91] uses different sub-models for partitioning (*gating models*) the input space as well as local prediction (*expert models*). In contrast to the decision tree the regions are not disjoint but there is a gradual change between regions. For each input point the predictions of the different experts are computed and used with weights determined by the gating network.

If we have m expert networks $z = f_j(\mathbf{x}, \theta_j)$, $j = 1, \ldots, m$, we need a gating network $g(\mathbf{x}, \phi)$ with one output $w_j = g_j(\mathbf{x}, \phi)$ for each expert network. To arrive at normalized weights these outputs are transformed by the 'softmax' function

$$\alpha_j(\mathbf{x}, \phi) = \frac{\exp(g_j(\mathbf{x}, \phi))}{\sum_{l=1}^{m} \exp(g_l(\mathbf{x}, \phi))} \tag{7}$$

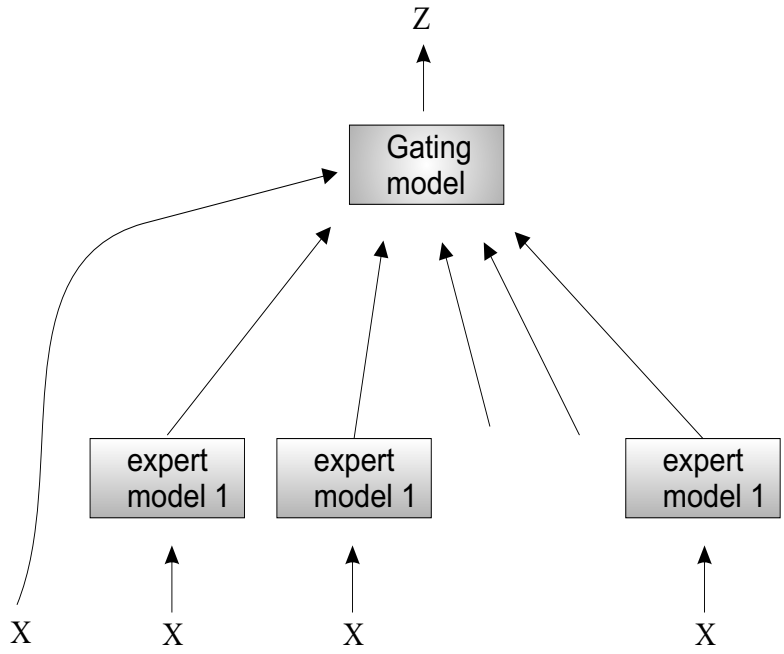

Fig. 1. In a mixture of experts a gating model defines probabilities for the different experts. The outputs of the expert models are weighted by these probabilities.

and the final output is the mixture of experts

$$z = \sum_{j=1}^{m} \alpha_j(\mathbf{x}, \phi)\, f_j(\mathbf{x}, \theta_j) + \varepsilon_j \qquad E(\varepsilon_j) = 0 \tag{8}$$

We may use virtually any model as expert model as long as it fits to the data $(z, \mathbf{x})$. Note that for Bayesian analysis a complete specification of the related distributions is required.

As an arbitrary number of experts may be combined we may use computationally simple models, whose combination may represent arbitrary complex dependencies. Candidates for continuous $z \in \Re$ are

- constants $z = c_j$. The gating network generates convex combinations of these constants.
- linear regression models $z = \sum_{i=1}^{k} x_i \theta_i + \varepsilon$ with normal error $\varepsilon \sim N(0, \sigma^2)$.
- quadratic or nonlinear regression models $z = \sum_{j=1}^{m} h_j(\theta_i) + \varepsilon$ with normal error $\varepsilon \sim N(0, \sigma^2)$ and fixed basis functions.
- Arbitrary generalized linear models [JPT97].

For discrete $z \in \{1, \ldots, r\}$ we may use any Bayesian classifier, and – in combination with the softmax function – arbitrary models with values in $\Re$. Simple examples are

- linear logistic model $f_j(\mathbf{x}, \theta) = \exp(\mathbf{x}'\theta_j) / \sum_{l=1}^{m} \exp(\mathbf{x}'\theta_l)$
- radial basis function models $f_j(\mathbf{x}, \theta) = h_j(\mathbf{x}, \theta_j, \sigma_j) / \sum_{l=1}^{m} h_l(\mathbf{x}, \theta_l, \sigma_l)$ with $h_j(\mathbf{x}, \theta_s, \sigma_s) = \prod_{j=1}^{k} \left(2\pi\sigma_j^2\right) \exp\left(-\frac{1}{2\sigma_j^2}\left(x_j - \theta_{sj}\right)^2\right)$

As gating network we may select any models $g_j(\mathbf{x}, \phi)$ with outputs in $\Re$ or any "probability model" $\alpha(\mathbf{x}, \phi)$ which generates a probability vector with m components, i.e. classifier models.

3.1 Prior Distributions

The choice of priors for a model is an important one in Bayesian inference. Priors embody the assumption about such aspects as the generative processes of the data and form of the model. The priors on a model are typically placed either on the structure (number of models) or the parameters of gate and expert models. The parameters of the gate and expert models are assumed to be mutually independent $p(\theta, \phi) = p(\theta)p(\phi)$. They may depend themselves on hyper-parameters, which themselves may be varied during the MCMC analysis. For the mean of radial basis functions as well as the means of regression models we use Normal priors with diagonal covariance matrix

$$p(\theta_s) = \prod_{j=1}^{k} \left(2\pi\rho_j^2\right) \exp\left(-\frac{1}{2\rho_j^2}\left(\theta_{sl} - \bar{\theta}_{sl}\right)^2\right) \tag{9}$$

For the variance σ^2 we use a Gamma prior on the inverse variance $\beta = 1/\sigma^2$

$$p(\log\beta) = \frac{1}{\Gamma(\tau)} \left(\frac{\beta}{v}\right)^{\tau} \exp\left(-\frac{\beta}{v}\right) \tag{10}$$

were $2/v$defines the prior sum of squared error that we might expect and 2τ defines the prior number of observations that we might expect an expert to see.

3.2 Comparison to other Models

It is instructive to visualize the regions defined by different types of experts. As shown in figure 2 logistic units $\exp(\mathbf{x}'\theta_j)$ put a "soft" threshold into the input space where they change their value from 0 to 1. Combined with the softmax function this results in mainly straight boundaries that partition the input space. It is important that each unit affects the whole partition. On the other hand radial basis function units $h_j(\mathbf{x}, \theta_j, \sigma_j)$ assign the region around the mean value θ_j to the corresponding unit. This leads to a Voronoi tesselation of the input space with linear boundaries between units, as long as the covariance terms for all units are identical.

Earlier Mixture of experts approaches therefore used logistic gating models but in a hierarchical fashion [JPT97], [Wat97]. In the bhighest layer two regions were defined, which were recursively partitioned by other gates of lower

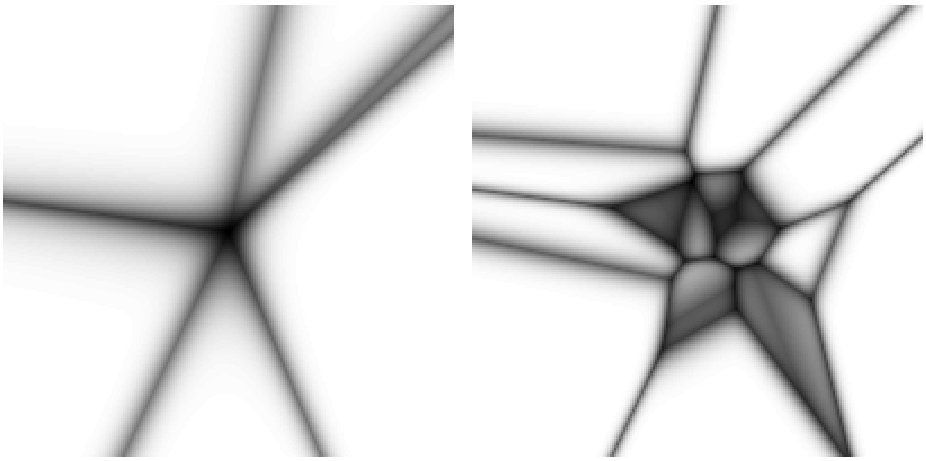

Fig. 2. Maximal probability of membership in a single region if the gates are logistic (left) or radial basis functions (right).

regions. For the Bayesian analysis these hierarchical mixtures of experts have a definite disadvantage: it is nearly impossible to change high-level gates in a MCMC analysis as this means that the whole tree of gates has to be deleted and rebuilt. Therefore Bayesian analyses tend to concentrate in a local minimum of the posterior density.

If we use non-hierarchical radial basis functions gates the changes only affect neighboring points. The MCMC algorithm can generate all plausible structures and effectively explore the posterior density. Therefore we prefer radial basis function units in our analysis.

There are a number of advanced statistical methods which may be applied to spatial problems in a similar way like the mixture of experts. They may be used in a semi-parametric fashion, i.e. they should be able to fit a wide set of functional relations in a nearly automatic way. They all can be evaluated in the framework of Bayesian statistics. This allows the flexible introduction of prior knowledge and the calculation of the uncertainty of statistical inference. Generalized additive models [VR97, p.281] and projection pursuit regression [FS82] define models on marginal variables and therefore are not able to fit arbitrary distributions. Local regression models are [CGJ95] are an attractive competitor of mixture of experts. Neural networks in the form of multilayer perceptrons [Nea95] also use logistic units and have the same convergence problems as hierarchical mixtures of experts. Similar problems occur for decision trees [PK98][CGM98] and multivariate adaptive regression splines (MARS) [Fri91][DMS97] which generate recursive partitions of the input space.

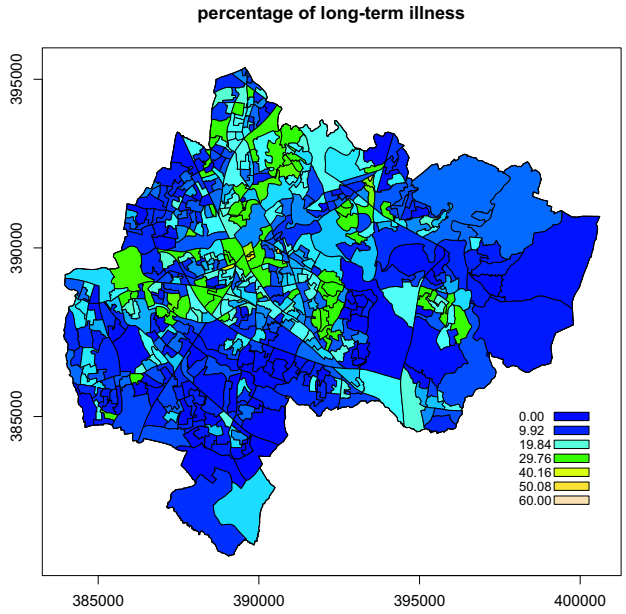

Fig. 3. Mean predicted value of long-term illness in Stockport using the Bayesian model.

4 Markov Chain Monte Carlo

The Markov Chain Monte Carlo analysis uses the following proposals to modify a model:

- Change the mean values of one gate unit.
- Change the variance of one variable for all gate units.
- Change the regression parameters of one expert model.
- Change the error variance of one expert model (not for classifier experts).
- Split one expert model into two (with different parameters).
- Merge two randomly selected expert models into one, whose parameters are the mean values of the components.

After an initial phase of several thousand iterations the MCMC algorithm reaches the stationary distribution of mixture of expert models. After this burn-in phase the models with all their parameters are stored for later use. We use the coda-package of R to determine the convergence to stationarity [BR98].

5 Application to Geodata

The mixture of experts model was implemented in the SPIN! system developed during the SPIN! project of the European community. It is a general tool for

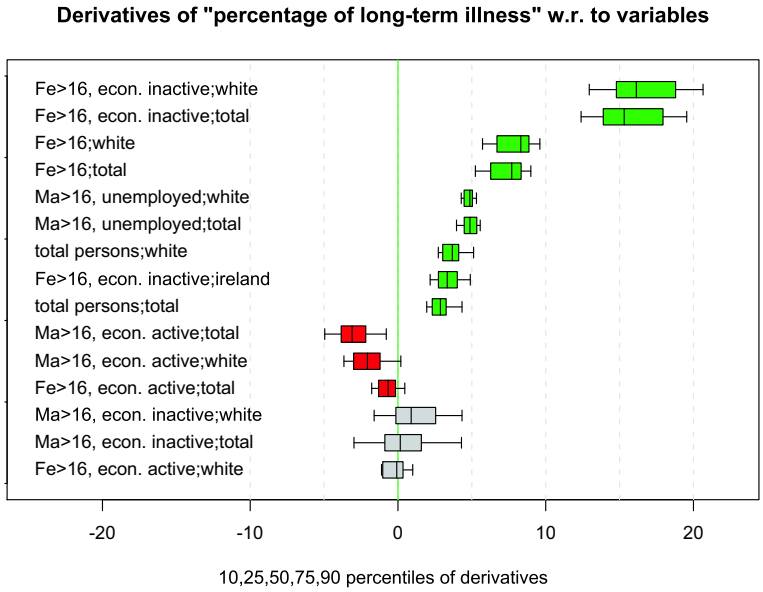

Fig. 4. Distributions of derivatives of long-term illness in Stockport for a specific ward. The boxes indicate 25% and 75% percentiles with the median in between. The outside "whiskers" are the 10% and 90% percentiles.

simulation Bayesian models by Markov Chain Monte Carlo. The system is implemented in Java to avoid compatibility problems.

As an introductory example we use data from Stockport, a town near Manchester, U.K. For small units of about 100 households (wards) we have statistics from the 1991 census including the basic demographic features as well as employment, car ownership, etc.

The Bayesian model was used to predict long-term illness from these figures. Hence the model is forced to adapt to the relation between the values of the input variables within the individual wards and the corresponding output variable variable long-term illness. The derivative of the output variable with respect to an input variable describes, how many units the output variable probably will increase if we increase the input variable for one unit. As the model is non-linear, the derivative will depend on the specific location, i.e. the input variables of the ward.

This figure may be important for planners if the want to check the stochastic relation between variables. It does not, however, imply, that the input variable actually may be changed, as many variables may not be controlled.

As our Bayesian model explicitly captures uncertainties the derivative is uncertain too. In figure 4 the resulting distribution of derivatives for a specific ward is shown. The graph can be generated interactively by clicking on a ward in the map above. The derivatives show that long-term illness in wards like the current

ward usually grows with the fraction of females aged higher than 16 years, which are economically inactive. This probably mainly applies to female pensioners. On the other hand long-term illness decreases if the number of economically active men increases.

On the workshop we will apply the approach to other data of North-West England.

References

[BR98] SP. Brooks and G.O. Roberts. Assessing the convergence of markov chain monte carlo algorithms. *Statistics and Computing*, pages 319–335, 1998.

[CGJ95] D. Cohn, Z. Ghahramani, and M. Jordan. Active learning with statistical models. In Tesauro et al. [TTL95], pages 705–712.

[CGM98] H. Chipman, E. George, and R. McCulloch. Bayesian CART model search. *JASA*, 93:935–960, 1998.

[DMS97] D. Denison, B. Mallick, and A. F. M. Smith. Bayesian mars. Technical report, Imperiag College, London, 1997.

[Fri91] Jerome H. Friedman. Adaptive spline networks. In R. Lippmann, J. Moody, and D. Touretzky, editors, *Advances in Neural Information Processing Systems 3*, volume 3, pages 675–683. Morgan Kaufmann Publishers, Inc., 1991.

[FS82] J. H. Friedman and W. Stuetzle. Projection pursuit methods for data analysis. *Modern Data Analysis*, pages 123–147, 1982.

[GCSR95] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman & Hall, London, 1995.

[Gre95] P. J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82:711–713, 1995.

[JJNH91] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3:79–87, 1991.

[JPT97] R. A. Jacobs, F. C. Peng, and M. A. Tanner. A Bayesian approach to model selection in hierarchical mixtures-of-experts architectures. *Neural Networks*, 10(2):231–241, 1997.

[Nea95] R. M. Neal. *Bayesian Learning for Neural Networks*. PhD thesis, Dep. of Computer Science, Univ. of Toronto, 1995.

[PK98] G. Paass and J. Kindermann. Bayesian classification trees with overlapping leaves applied to credit scoring. In X. Wu, R. Kotagiri, and K. B. Korb, editors, *Research and Development in Knowledge Discovery and Data Mining*, pages 234–245. Springer Verlag, 1998.

[Tie94] L. Tierney. Markov chains for exploring posterior distributions. *Annals of Statistics*, 22:1701–1762, 1994.

[TTL95] G. Tesauro, D. Touretzky, and T. Leen, editors. *Advances in Neural Information Processing Systems 7*. The MIT Press, 1995.

[VR97] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S-Plus*. Springer Verlag, New York, 3rd edition, 1997.

[Wat97] S. R Waterhouse. *Classification and Regression using Mixtures of Experts*. PhD thesis, Cambridge University Engineering Dept., October 1997.