Regression density estimation with variational methods and stochastic approximation

David J. Nott, Siew Li Tan, Mattias Villani and Robert Kohn

Faculty of Science
Department of Statistics and Applied Probability

Introduction

Regression density estimation is the problem of flexibly estimating a response distribution as a function of covariates. Our approach considers flexible mixtures of heteroscedastic experts (MHE) regression models where the response distribution is a normal mixture, with the component means, variances and mixture weights all varying as a function of covariates. Fast variational approximation (VA) methods are developed for inference as computationally intensive MCMC methods are difficult to apply when it is desired to fit models repeatedly. A variational approximation for MHE models is described where the variational lower bound is in closed form and this basic approximation can be improved by using stochastic approximation (SA) methods. The advantages of our approach for model choice and evaluation compared to MCMC based approaches are illustrated.

Heteroscedastic mixtures of experts model

Observed responses $y_1, ..., y_n$ are modelled by a mixture of experts model [1] with k mixture components. Conditional on latent variable, $\delta_i = j$,

$$y_i|\beta, \alpha \sim N(x_i^T\beta_j, \exp(\alpha_j^T z_i))$$

where x_i, z_i are vectors of covariates and β_j, α_j are vectors of unknown parameters. The prior for δ_i is

$$P(\delta_i = j|\gamma) = p_{ij} = \frac{\exp(\gamma_j^T v_i)}{\sum_{l=1}^k \exp(\gamma_l^T v_i)}, \quad j = 1, ..., k$$

where v_i is a vector of covariates and γ_j, $j = 2, ..., k$ are vectors of unknown parameters, with $\gamma_1 \equiv 0$ for identifiability. For Bayesian inference, assume independent priors, $\beta_j \sim N(\mu^0_{\beta j}, \Sigma^0_{\beta j})$, $\alpha_j \sim N(\mu^0_{\alpha j}, \Sigma^0_{\alpha j})$, $j = 1, ..., k$, and $\gamma = (\gamma_2^T, ..., \gamma_k^T)^T \sim N(\mu^0_\gamma, \Sigma^0_\gamma)$.

Variational approximation

Consider a VA $q(\theta)$ to the joint posterior of all parameters θ. A parametric form is chosen for $q(\theta)$ and we try to minimize the Kullback-Leibler divergence between $q(\theta)$ and $p(\theta|y)$. This is equivalent to maximizing the lower bound on the log marginal likelihood. Let $q(\theta) = q(\delta)q(\beta)q(\alpha)q(\gamma)$,

$$q(\delta) = \prod_{i=1}^n q(\delta_i), \quad q(\beta) = \prod_{i=1}^k q(\beta_i) \quad q(\alpha) = \prod_{i=1}^k q(\alpha_i)$$

where $q(\beta_i)$ is $N(\mu^q_{\beta i}, \Sigma^q_{\beta i})$, $q(\alpha_i)$ is $N(\mu^q_{\alpha i}, \Sigma^q_{\alpha i})$, $q(\delta_i = j) = q_{ij}$ and $q(\gamma) = I_{\gamma = \mu^q_\gamma}$. We derived a closed form of the variational lower bound and optimized it with respect to each set of variational parameters with others held fixed in a gradient ascent algorithm. Approximate methods were developed for dealing with variance parameters whose updates cannot be obtained in closed form. To initialize the algorithm, we generate a initial clustering of the data and set $\mu^q_{\alpha j} = \Sigma^q_{\alpha j} = 0$ for $j = 1, ..., k$ and q_{ij} as 1 if the ith observation lies in cluster j, 0 otherwise. Parameter updates are performed iteratively until the increase in the lower bound is less than a tolerance. To deal with multiple modes, we considered 20 random clusterings, performed short runs of the algorithm and follow only the solution with the highest lower bound to convergence.

Cross-validation

We use the log predictive density score (LPDS) to measure predictive performance. In B-fold cross-validation (CV), the data is split randomly into B equal parts. A training set T_i is formed by leaving out one of the B parts (say, F_i) from the complete data set.

$$\text{LPDS} = \frac{1}{B}\sum_{i=1}^B \log p(y_{F_i}|X_{F_i}, y_{T_i}),$$

As $\log p(y_F|X_F, y_T) = \log \int p(y_F|X_F, \theta)p(\theta|y_T)d\theta$, we can estimate the LPDS by replacing $p(\theta|y)$ with $q(\theta)$.

Model choice in time series

Predictive performance is measured by

$$\log p(y_{>T}|y_{\leq T}) = \sum_{i=1}^{T^*} \log p(y_{T+i}|y_{\leq T+i-1}).$$

Since $p(y_{T+i}|y_{\leq T+i-1}) = \int p(y_{T+i}|\theta, y_{\leq T+i-1})p(\theta|y_{\leq T+i-1})d\theta$, different posterior distributions are involved as successive points are added to the observed data. Our variational approach is very efficient for implementing such sequential updating as result of the variational optimization from the last time step can be used to initialize optimization for the current time step so that the convergence time of the variational scheme is small.

Improving the basic approximation

Variational approximations can underestimate the variance of the posterior. We propose to improve estimates obtained from VA by using SA which has reduced computational cost compared to MCMC. [2] independently proposed a similar approach but we offer improvements in the form of an improved gradient estimate in the SA procedure, and the idea of perturbing only the mean and scale of an initial VA.

Emulation of a rainfall runoff model

Our goal is to emulate the streamflow response (y) of the Australian Water Balance Model (Boughton, 2004) as a function of maximum storage capacity (x_1) and baseflow recession factor (x_2). The data consists of model simulations for 500 parameters values and we considered fitting five models. Models A, B, C and D are MHE models with both predictors in the mean and variance models, having respectively 2, 3, 4 and 5 mixture components. Model E is similar to model C but with only an intercept in the variance model (homoscedastic mixture of experts).

		Model A	Model B	Model C	Model D	Model E
ML	VA	-803.4	-688.4	-678.5	-682.8	-729
LPDS	VA	-65.9	-54.5	-51.5	-52.1	-57.2
LPDS	MCMC	-65.5	-54.2	-51.2	-51.4	-57.4

Table 1: Marginal log likelihood estimated by variational lower bound (1st row) and LPDS with ten-fold CV estimated by VA (2nd row) and MCMC (3rd row).

LPDS values in Table 1 computed by VA compare well with those obtained by MCMC. The results suggest that model C with 4 mixture components is adequate and Figure 1 summarizes model C. Here observations are separated into clusters according to which mixture component each observation is most likely to belong to.

		Model A	Model B	Model C	Model D	Model E
Full	VA	88	146	215	274	254
data	MCMC	330	473	650	825	659
CV	VA	121	184	281	393	276
	MCMC	2941	4409	5979	7626	5929

Table 2: CPU times (in seconds) for full data and CV computations by VA and MCMC.

CV computation times in Table 2 indicate an approximately 20 fold speed up for all models by using VA when using just 10,000 iterations in the MCMC sampling. The difficulties of convergence assessment in the MCMC approach are also avoided by the variational method.

For model C, we compared posterior distributions obtained via MCMC with both our simple VA and VA incorporating SA correction. Computation of the SA correction took 166 seconds of CPU time. The SA correction is helpful for obtaining an improved approximation for at least some of the parameters, with the estimated posterior marginals from SA generally being closer to the Monte Carlo estimated marginals than the simple variational estimated marginals.

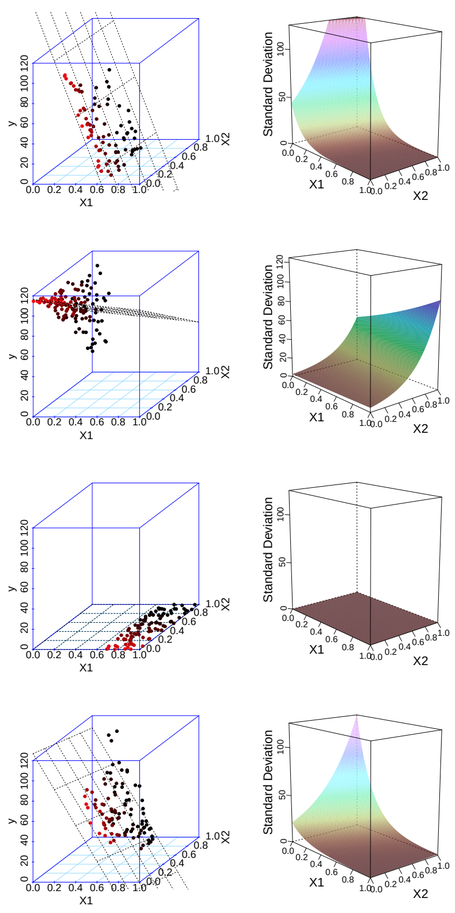

Figure 1: Fitted component means (1st column) and standard deviations (2nd column) for model C.

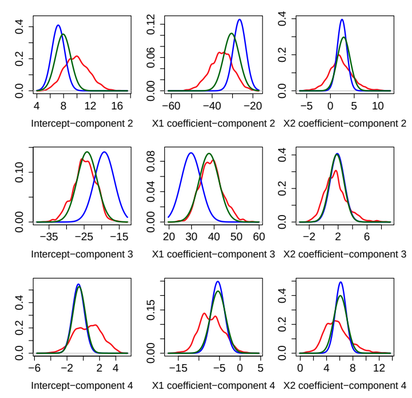

Figure 2: Marginal posterior distributions for parameters in the gating function estimated by Monte Carlo method (red), simple VA (blue) and VA with SA correction (green) for model C.

Time series example

Following [3], we consider data for the S&P500 stock market index, taking 4646 daily returns from Jan 1, 1990 to May 29, 2008 as training set and the subsequent 199 daily returns from May 30, 2008 to March 13, 2009 as validation set. Our response y_t is $\log p_t/p_{t-1}$ where p_t is the closing S&P500 index on day t. We consider MHE models with only an intercept term in the mean model but an intercept and the covariates `LastWeek`, `LastMonth` and `MaxMin95` in the variance model and gating function and $m = 1, 2, 3$ and 4 experts.

	No. of mixture components			
	1	2	3	4
No sequential updating, MCMC	-477.8	-471.2	-469.0	-470.6
No sequential updating, VA	-478.0	-470.1	-470.1	-471.7
With sequential updating, VA	-477.7	-470.0	-470.1	-473.3

Table 3: LPDS values and MCMC method with approximation of [3] (1st line), variational method with approximation of [3] (2nd line) and variational method with sequential updating (last line).

	No. of mixture components			
	1	2	3	4
Initial fit MCMC	504	2463	3427	4417
Initial fit VA	1	739	1022	1442
Initial fit + validation VA	250	1902	2552	4754

Table 4: Computation times (seconds) for LPDS calculations. Rows 1-3 respectively are times for initial fit for MCMC, initial fit for VA, and initial fit plus sequential updating for validation for VA.

[3] uses an approximation posterior that is not updated after end of the training period. From Table 3, based on the largest LPDS, a two component mixture is adequate. Computation times for MCMC and VA in Table 4 indicate a roughly 200 fold speed up from employing the variational method as computational cost for the initial fit for MCMC needs to be multiplied by approximately 199 to get total computational cost.

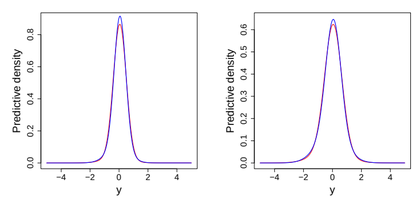

Figure 3: Estimated predictive densities at covariate values for $t = 1000$ (left) and $t = 4000$ (right) based on entire training data set using MCMC (red) and VA (blue) for 2 component mixture model.

The MCMC and variational predictive densities in Figure 3 are nearly indistinguishable, so that the VA provides excellent predictive inference here.

References

[1] Jacobs, R., Jordan, M., Nowlan, S. and Hinton, G. (1991). Adaptive mixtures of local experts. *Neural Computation*, 3, 79–87.

[2] Ji, C., Shen, H. and West, M. (2010). Bounded approximations for marginal likelihoods. Technical report, Duke University ISDS, available at http://ftp.stat.duke.edu/WorkingPapers/10-05.html

[3] Li, F., Villani, M., and Kohn, R. (2009). Flexible modeling of conditional distributions using smooth mixtures of asymmetric student t densities. Riksbank Research Paper Series No. 64.

Acknowledgements

Siew Li Tan and David Nott gratefully acknowledge support of the Singapore-Delft Water Alliance (SDWA).