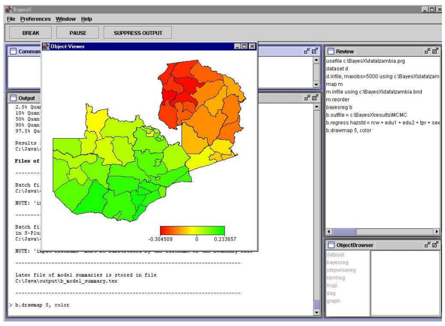
BayesX: Analysing Bayesian structured additive regression models

Andreas Brezger[1], Thomas Kneib[1] and Stefan Lang[2]

[1] Department of Statistics, University of Munich

[2] Institute of Empirical Economic Research, University of Leipzig

What is BayesX?



- The software BayesX provides powerful regression tools for analysing structured additive regression (STAR) and survival models. STAR models cover a number of well known model classes as special cases, e.g. generalised additive models, generalised additive mixed models, geoadditive models, dynamic models, varying coefficient models, and geographically weighted regression.
- BayesX supports both full Bayesian inference based on Markov chain Monte Carlo simulation techniques and empirical Bayes inference based on a mixed model representation of STAR models.
- In its current form, BayesX runs only under the various versions of the Windows operating system. A Linux version and an interface to R are work in progress.
- Contributions by Christiane Belitz, Eva–Maria Fronk, Andrea Hennerfeind, Manuela Hummel, Alexander Jerak, Petra Kragler and Leyre Osuna Echavarría.

Bayesian structured additive regression

- Structured additive regression extends and unifies several additive and geoadditive regression approaches.
- For exponential family models, a structured additive predictor is of the form

$$\eta_i = f_1(\nu_{i1}) + \ldots + f_p(\nu_{ip}) + u_i'\gamma,$$

where the ν_j are generic covariates of different type and dimension, and the f_j are (not necessarily smooth) functions of the covariates.

- Supported model terms include:
  - Penalised splines and random walk priors for nonparametric effects $f(x)$ of continuous covariates x or time trends $f(t)$.
  - Bivariate tensor product penalised splines for interaction surfaces $f(x, z)$.
  - Varying coefficient terms with continuous and spatial effect modifiers.
  - State space models for time-varying seasonal patterns $f(t)$.
  - Random intercepts and random slopes.
  - (Intrinsic) Markov random field priors for spatial effects $f_{spat}(s)$ based on regional data $s \in \{1, \ldots, S\}$.
  - Stationary Gaussian random field priors for spatial effects $f_{spat}(s)$ based on point-referenced data $s = (s_x, s_y)$.
- Supported univariate response distributions:
  - Gaussian responses with identity link.
  - Binary responses with logit, probit and complementary log-log link.
  - Poisson responses with log-link.
  - Gamma responses with log-link.
  - Negative Binomial responses with log-link.
- Extensions for categorical regression models:
  - Multinomial logit (and probit) models for unordered responses:

$$P(Y = r) = \frac{\exp(\eta^{(r)})}{1 + \sum_{s=1}^{k-1} \exp(\eta^{(s)})}.$$

$$\eta^{(r)} = u'\alpha^{(r)} + \bar{w}^{(r)\prime}\delta + \sum_{j=1}^{l} f_j^{(r)}(\nu_j) + \sum_{j=l+1}^{p} \bar{f}_j(\nu_j^{(r)}).$$

  - Global covariates with category-specific effects.
  - Category-specific covariates with global effects.
  - Cumulative and sequential logit and probit models for ordered responses:

$$\eta^{(r)} = \theta^{(r)} - u'\alpha - w'\delta^{(r)} - \sum_{j=1}^{l} f_j(\nu_j) - \sum_{j=l+1}^{p} f_j^{(r)}(\nu_j).$$

  - Global effects and category-specific effects.
  - In ordinal models, category-specific effects induce complicated constraints.
- Extensions for the analysis of survival times:
  - Joint estimation of covariate effects and baseline hazard rate.
  - Time-varying effects $g(t)x$ of covariates x.
  - Inclusion of (piecewise constant) time-varying covariates.
  - Left, right and interval censoring as well as left truncation.

Inference

- All effects can be subsumed in a unified framework.
- Each model term is associated with a vector of regression coefficients ξ_j with a multivariate Gaussian, partially improper prior

$$p(\xi_j|\tau_j^2) \propto \exp\left(-\frac{1}{2\tau_j^2}\xi_j' K_j \xi_j\right).$$

- The precision matrix K_j acts as a penalty matrix and the variance τ_j^2 represents the smoothing parameter.
- Inference can be performed either fully Bayesian based on MCMC or empirically Bayesian based on a mixed model representation.
- Key features of fully Bayesian inference:
  - Additional inverse Gamma priors are assigned to the variance parameters τ_j^2.
  - The precision matrix K_j is usually sparse. Sparse matrix computations can be used for the construction of efficient updating schemes.
  - Update parameters of one term in a block based on iteratively weighted least squares proposals derived from generalised additive model backfitting equations.
  - Data augmentation in models with underlying latent Gaussian variables allows for Gibbs sampling in binary and categorical probit models.
- Key features of empirical Bayes inference:
  - Consider variance parameters as unknown constants and estimate them based on their marginal posterior.
  - Mixed model methodology can be used to do this: Split the partially improper prior for ξ_j in two parts with proper and improper prior density. The resulting model can be interpreted as a mixed model.
  - Marginal likelihood / Restricted maximum likelihood estimates for the variances can be derived from standard mixed model algorithms. This usually involves a Laplace approximation to the marginal posterior.

Example: Leukemia survival data

- Survival time of adults after diagnosis of acute myeloid leukemia.
- 1,043 cases diagnosed between 1982 and 1998 in Northwest England.
- 16 % (right) censored.
- Continuous and categorical covariates:

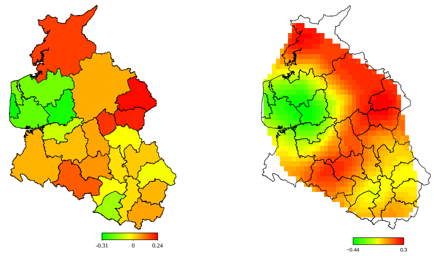
age	age at diagnosis,
wbc	white blood cell count at diagnosis,
sex	sex of the patient,
tpi	Townsend deprivation index.

- Spatial information in different resolution.
- Suitable model for the hazard rate:

$$\lambda(t; \cdot) = \exp[f_0(t) + f_1(age) + f_2(wbc) + f_3(tpi) + f_{spat}(s_i) + \gamma_1 sex]$$

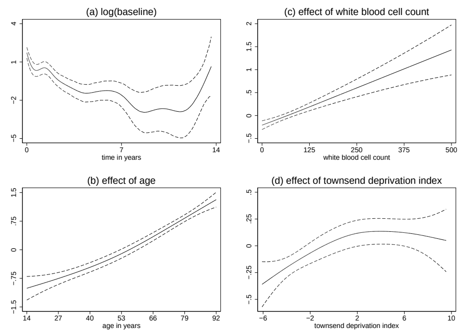
where

  - $f_0(t) = \log(\lambda_0(t))$ is the log-baseline-hazard,
  - f_1, f_2, f_3 are nonparametric functions of age, white blood cell count and deprivation, and
  - f_{spat} is a spatial function.
- Results for the spatial effect:



District-level analysis Individual-level analysis

- Results for the nonparametric effects:



Further information

- Download and further information:

`http://www.stat.uni-muenchen.de/ bayesx`

- Contact:

`bayesx@stat.uni-muenchen.de`

- Selected references:

  - Brezger, A. & Lang, S. (2006). Generalized structured additive regression based on Bayesian P-splines. *Computational Statistics and Data Analysis*, **50**, 967-991.
  - Brezger, A., Kneib, T. & Lang, S. (2005). BayesX: Analyzing Bayesian structured additive regression models. *Journal of Statistical Software*, **14** (11).
  - Fahrmeir, L., Kneib, T. & Lang, S. (2004). Penalized structured additive regression for space-time data: a Bayesian perspective. *Statistica Sinica* **14**, 731–761.
  - Hennerfeind, A., Brezger, A. & Fahrmeir, L. (2006). Geoadditive survival models. *Journal of the American Statistical Association*, to appear.
  - Kneib, T. & Fahrmeir, L. (2006). Structured additive regression for categorical space-time data: A mixed model approach. *Biometrics*, to appear.
  - Kneib, T. & Fahrmeir, L. (2006). A mixed model approach for geoadditive hazard regression. *Scandinavian Journal of Statistics*, to appear.