A.1.2 Generalized additive models (GAM's)

Model description A very useful generalization of the ordinary multiple regression

$$y_i = \mu + \beta_1 x_{1,i} + \cdots + \beta_p x_{p,i} + \varepsilon_i,$$

is the class of additive models,

$$y_i = \mu + f_1(x_{1,i}) + \cdots + f_p(x_{p,i}) + \varepsilon_i. \tag{A.1}$$

Here, the f_j are 'nonparametric' components which can be modelled by penalized splines. When this generalization is carried over to generalized linear models, and we arrive at the class of GAM's (Hastie & Tibshirani 1990). From a computational perspective penalized splines are equivalent to random effects, and thus GAM's fall naturally into the domain of ADMB-RE.

For each component f_j in (A.1) we construct a design matrix $\mathbf{X}$ such that $f_j(x_{i,j}) = \mathbf{X}^{(i)}\mathbf{u}$, where $\mathbf{X}^{(i)}$ is the ith row of $\mathbf{X}$ and $\mathbf{u}$ is a coeffisient vector. We use the R-function `splineDesign` (from the `splines` library) to construct a design matrix $\mathbf{X}$. To avoid overfitting we add a first order difference penalty (Eilers & Marx 1996) :

$$-\lambda^2 \sum_{k=2} (u_k - u_{k-1})^2 , \tag{A.2}$$

to the ordinary GLM loglikelihood, where λ is a smoothing parameter to be estimated. By viewing $\mathbf{u}$ as a random effects vector with the above Gaussian prior, and by taking λ as a hyper-parameter, it becomes clear that GAM's are naturally handled in ADMB-RE.

Implementation details

- A computationally more efficient implementation is obtained by moving λ from the penalty term to the design matrix, i.e. $f_j(x_{i,j}) = \lambda^{-1}\mathbf{X}^{(i)}\mathbf{u}$.

- Since (A.2) does not penalize the mean of $\mathbf{u}$, we impose the restriction that $\sum_{k=1} u_k = 0$ (see the `union.tpl` for details). Without this restriction the model would be over-parameterized since we allready have an overall mean μ in (A.1).

- To speed up computations the parameter μ (and other regression parameters) should be given 'phase 1' in ADMB, while the λ's and the $\mathbf{u}$'s should be given given 'phase 2'.

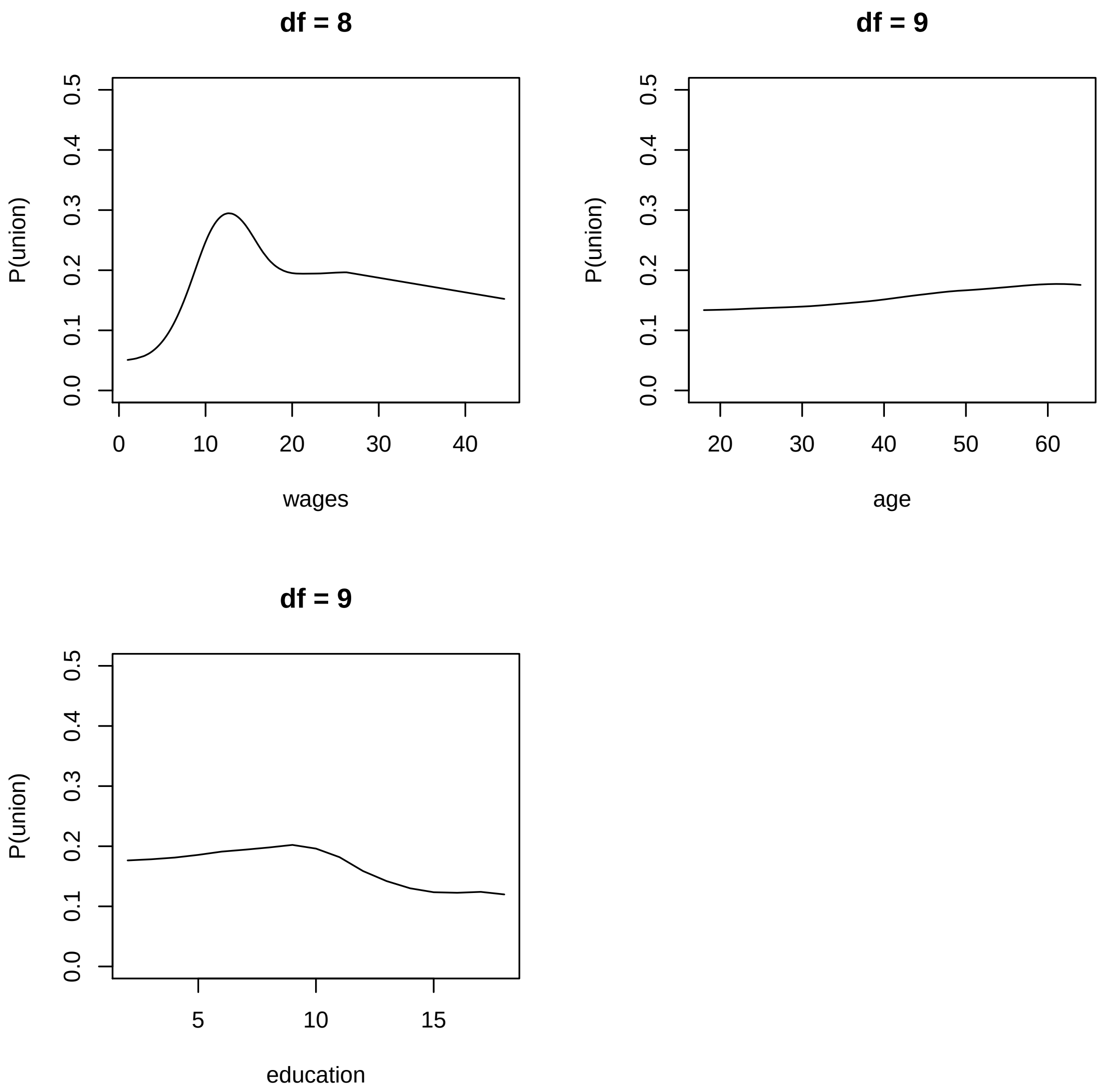

Figure A.1: Probablity of membership as a function of covariates. In each plot, the remaining covariates are fixed at their sample means. The effective degrees of freedom (df) are also given (Hastie & Tibshirani 1990).

The Wage-union data The data, which are available from Statlib (`lib.stat.cmu.edu/`), contain information for each of 534 workers about whether they are members ($y_i = 1$) of a workers union or not ($y_i = 0$). We study the probability of membership as a function of six covariates. Expressed in the notation used by the R (S-Plus) function `gam` the model is:

```
union ~race + sex + south + s(wage) + s(age) + s(ed), family=binomial
```

Here, `s()` denotes a spline functions with 20 knots each. For `wage` a cubic spline is used, while for `age` and `ed` quadratic splines are used. The total number of random effects that arrise from the three corresponding **u** vectors is 64. Figure A.1 shows the estimated nonparametric components of the model. The time taken to fit the model was 165 seconds.

Extentions

- The linear predictor may be a mix of ordinary regression terms ($f_j(x) = \beta_j x$) and nonparametric terms. ADMB-RE offers a unified approach to fitting such models, in which the smoothing parameters λ_j and the regression parameters β_j are estimated simultaneously.
- It is straight forward in ADMB-RE to add 'ordinary' random effects to the model, for instance to accomodate for correlation within groups of observations, as in Lin & Zhang (1999).

Files http://otter-rsch.com/admbre/examples/union/union.html

Bibliography

ADMB Development Core Team (2009), *An Introduction to AD Model Builder*, ADMB project.

ADMB Foundation (2009), 'ADMB-IDE: Easy and efficient user interface', *ADMB Foundation Newsletter* **1**, 1–2.

Eilers, P. & Marx, B. (1996), 'Flexible smoothing with B-splines and penalties', *Statistical Science* **89**, 89–121.

Harvey, A., Ruiz, E. & Shephard, N. (1994), 'Multivariate stochastic variance models', *Review of Economic Studies* **61**, 247–264.

Hastie, T. & Tibshirani, R. (1990), *Generalized Additive Models*, Vol. 43 of *Monographs on Statistics and Applied Probability*, Chapman & Hall, London.

Kuk, A. Y. C. & Cheng, Y. W. (1999), 'Pointwise and functional approximations in Monte Carlo maximum likelihood estimation', *Statistics and Computing* **9**, 91–99.

Lin, X. & Zhang, D. (1999), 'Inference in generalized additive mixed models by using smoothing splines', *J. Roy. Statist. Soc. Ser. B* **61**(2), 381–400.

Pinheiro, J. C. & Bates, D. M. (2000), *Mixed-Effects Models in S and S-PLUS*, Statistics and Computing, Springer.

Rue, H. & Held, L. (2005), *Gaussian Markov random fields: theory and applications*, Chapman & Hall/CRC.

Ruppert, D., Wand, M. & Carroll, R. (2003), *Semiparametric Regression*, Cambridge University Press.

Skaug, H. & Fournier, D. (2006), 'Automatic approximation of the marginal likelihood in non-gaussian hierarchical models', *Computational Statistics & Data Analysis* **56**, 699–709.

Zeger, S. L. (1988), 'A regression-model for time-series of counts', *Biometrika* **75**, 621–629.