Some Results on Penalized Spline Estimation in Generalized Additive and Semiparametric Models

Gerda Claeskens
Limburgs Universitair Centrum, Center for Statistics
Universitaire Campus
B-3590 Diepenbeek, Belgium
Gerda.Claeskens@luc.ac.be

Marc Aerts
Limburgs Universitair Centrum, Center for Statistics
Universitaire Campus
B-3590 Diepenbeek, Belgium

M. P. Wand
Harvard School of Public Health, Department of Biostatistics
Huntington Avenue
Boston, Massachusetts 02115, USA

1. Introduction

In spline regression models, the smoothness of the fitted model depends on the knots via their location and how many knots there are. An alternative to knot selection is to keep all the knots, but to restrict their influence on the fitted model, that is, to constrain the values of these coefficients which correspond to the spline basis functions. For example, we can bound the squared L_2 norm by some constant. In this case, for a classical additive regression model with d covariates, the fitted values are given by $\widehat{\boldsymbol{f}}_{\boldsymbol{\alpha}} = \boldsymbol{G}_{\boldsymbol{\alpha}}\boldsymbol{Y}$ where $\boldsymbol{G}_{\boldsymbol{\alpha}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{A}_{\boldsymbol{\alpha}})^{-1}\boldsymbol{X}^T$, $\boldsymbol{A}_{\boldsymbol{\alpha}} = \text{blockdiag}_{1\le j\le d}(\alpha_j\boldsymbol{D}_j)$, $\boldsymbol{D}_j = \text{diag}(\boldsymbol{0}_{p_j\times 1}, \boldsymbol{1}_{K_j})$, $\boldsymbol{X}$ is the full design matrix of the covariates (up to polynomial degree p_j) and the spline basis functions $\{(x_{ji} - \kappa_{jk})_+\}^{p_j}$, where $\kappa_{j1}, \ldots, \kappa_{jK_j}$ is a set of knots for the jth covariate. Since the amount of smoothing is determined by $\boldsymbol{\alpha}$, this is the vector of smoothing parameters. See also Marx and Eilers (1998). Some other penalty functions are proposed by Ruppert and Carroll (1997).

We obtain closed form approximations to the bias and variance of the estimators, not just for additive models, but for generalized additive models (GAM). The results can for example be used to provide rough starting values for the smoothing parameters and they provide some backup for commonly used degrees of freedom approximations.

2. Approximation of the risk and degrees of freedom

In a GAM the parameter of interest is modeled as an additive function of the covariates. We will model $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}$ with $\boldsymbol{X}$ the spline design matrix. For $\boldsymbol{U}_\eta$ the vector of first partial derivatives and $\boldsymbol{J}_\eta$ the matrix of minus second partial derivatives of the log likelihood with respect to $\boldsymbol{\eta}$, we obtain the estimators $\widehat{\boldsymbol{\beta}}$ via iteratively reweighted ridge regression, where the adjusted dependent variable is defined as $\boldsymbol{Z}_\eta = \boldsymbol{\eta} + \boldsymbol{J}_\eta^{-1}\boldsymbol{U}_\eta$. A first order approximation of the estimator of the coefficient $\boldsymbol{\beta}$ is given by $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{J}_\eta\boldsymbol{X} + \boldsymbol{A})^{-1}\boldsymbol{X}^T\boldsymbol{J}_\eta\boldsymbol{Z}_\eta$, which allows us to extend the asymptotic results of Wand (1998) in a one-covariate regression model to generalized additive models.

The overall risk is measured by $\text{MASE}(\widehat{\boldsymbol{\eta}}) = \frac{1}{n}\text{tr}\,\{\boldsymbol{G}\text{Var}(\boldsymbol{U}_\eta)\boldsymbol{G}\} + \frac{1}{n}\|\boldsymbol{G}\boldsymbol{F}\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\eta}\|^2$, where $\boldsymbol{G} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{F}\boldsymbol{X} + \boldsymbol{A}_{\boldsymbol{\alpha}})^{-1}\boldsymbol{X}^T$, and $\boldsymbol{F} = E[\boldsymbol{J}_\eta]$. An asymptotic approximation to MASE yields an expression for the AMASE-optimal smoothing parameters.

We follow Hastie and Tibshirani (1990) in defining the degrees of freedom for the jth component in a generalized additive model as $\text{df}_j = \text{tr}(\boldsymbol{G}_j\boldsymbol{F})$ where $\boldsymbol{G}_j = [\mathbf{0}\cdots\mathbf{0}\ \boldsymbol{X}_j\ \mathbf{0}\cdots\mathbf{0}]\times(\boldsymbol{X}^T\boldsymbol{F}\boldsymbol{X} + \boldsymbol{A}_{\boldsymbol{\alpha}})^{-1}\boldsymbol{X}^T$. We can show that

$$\frac{\text{tr}(\boldsymbol{G}_j\boldsymbol{F})}{\text{tr}(\boldsymbol{S}_j\boldsymbol{F})} - 1 \approx -\alpha_j\,\frac{\text{tr}[\boldsymbol{X}_{[-j]}\{\boldsymbol{X}_{[-j]}^T\boldsymbol{F}(\boldsymbol{I} - \boldsymbol{S}_{j0}\boldsymbol{F})\boldsymbol{X}_{[-j]}\}^{-1}\boldsymbol{X}_{[-j]}^T\boldsymbol{F}\boldsymbol{B}_j\boldsymbol{F}]}{p_j + K_j - \alpha_j\text{tr}\{(\boldsymbol{X}_j^T\boldsymbol{F}\boldsymbol{X}_j)^{-1}\boldsymbol{D}_j\}},$$

where $\boldsymbol{X}_{[-j]}$ is the design matrix of all components but the jth one, $\boldsymbol{B}_j = \boldsymbol{X}_j(\boldsymbol{X}_j^T\boldsymbol{F}\boldsymbol{X}_j)^{-1}\boldsymbol{D}_j\times(\boldsymbol{X}_j^T\boldsymbol{F}\boldsymbol{X}_j)^{-1}\boldsymbol{X}_j^T$ and $\boldsymbol{S}_j = \boldsymbol{X}_j(\boldsymbol{X}_j^T\boldsymbol{F}\boldsymbol{X}_j + \alpha_j\boldsymbol{D}_j)^{-1}\boldsymbol{X}_j^T$. The degree of freedom value for the jth component might be approximated by $\text{tr}(\boldsymbol{S}_j\boldsymbol{F})$. This has the computational advantage that only that part of the design matrix related to the jth covariate needs to be used.

In semiparametric models we can use the same methodology as in fully nonparametric models. If the design matrix is partitioned as $\boldsymbol{X} = [\boldsymbol{X}_{\text{parm}},\ \boldsymbol{X}_{\text{nonp}}]$, where $\boldsymbol{X}_{\text{parm}} = [\boldsymbol{X}_1, \cdots, \boldsymbol{X}_q]$ and $\boldsymbol{X}_{\text{nonp}} = [\boldsymbol{X}_{q+1}, \cdots, \boldsymbol{X}_d]$, the smoother matrix $\boldsymbol{G} = \boldsymbol{X}_{\text{parm}}(\boldsymbol{X}_{\text{parm}}^T\boldsymbol{X}_{\text{parm}})^{-1}\boldsymbol{X}_{\text{parm}}^T + \boldsymbol{R}(\boldsymbol{R}^T\boldsymbol{R} + \boldsymbol{A}_{\text{nonp}})^{-1}\boldsymbol{R}^T$ where $\boldsymbol{R} = (\boldsymbol{I} - \boldsymbol{X}_{\text{parm}}(\boldsymbol{X}_{\text{parm}}^T\boldsymbol{X}_{\text{parm}})^{-1}\boldsymbol{X}_{\text{parm}}^T)\boldsymbol{X}_{\text{nonp}}$ and $\boldsymbol{A}_{\text{nonp}} = \text{diag}_{q+1\le j\le d}(\alpha_j\boldsymbol{D}_j)$. An optimal smoothing parameter can be obtained by minimizing the MASE with respect to $\boldsymbol{\alpha}$. This boils down to replacing $\boldsymbol{X}$ by $\boldsymbol{R}$ in the previous expression. Another strategy is to select the smoothing parameters optimally for estimation of the nonparametric part only. More details and results are given in Aerts, Claeskens and Wand (1999).

REFERENCES

Aerts, M., Claeskens, G. and Wand, M.P. (1999). Some theory for penalized spline additive models. Submitted.

Hastie, T.J. and Tibshirani, R.J. (1990). Generalized additive models. Chapman & Hall. London.

Marx, B.D. and Eilers, P.H.C. (1998). Direct generalized additive modeling with penalized likelihood. Journal of Computational Statistics and Data Analysis, 28, 193–209.

Ruppert, D. and Carroll, R.J. (1997). Penalized regression splines. Unpublished manuscript.

Wand, M.P. (1999). On the optimal amount of smoothing in penalized spline regression. Biometrika, to appear.

RÉSUMÉ

On obtient une approximation du biais et de la variance des estimateurs d'une régression pénalisée utilisant des splines, pour des modèles additifs généralisés. Les résultats peuvent, par exemple, être utilisés pour obtenir des valeurs initiales pour les paramètres de lissage et fournir un support additionnel pour les approximations par les degrés de liberté habituellement utilisées.