Jerak, Alexander; Wagner, Stefan

Working Paper

Modeling Probabilities of Patent Oppositions in a Bayesian Semiparametric Regression Framework

Discussion Paper, No. 323

Provided in Cooperation with:
Collaborative Research Center (SFB) 386: Statistical Analysis of discrete structures - Applications in Biometrics and Econometrics, University of Munich (LMU)

Suggested Citation: Jerak, Alexander; Wagner, Stefan (2003) : Modeling Probabilities of Patent Oppositions in a Bayesian Semiparametric Regression Framework, Discussion Paper, No. 323, Ludwig-Maximilians-Universität München, Sonderforschungsbereich 386 - Statistische Analyse diskreter Strukturen, München,
https://doi.org/10.5282/ubm/epub.1704

This Version is available at:
https://hdl.handle.net/10419/23881

Modeling Probabilities of Patent Oppositions in a Bayesian Semiparametric Regression Framework

Alexander Jerak*, Stefan Wagner**

* Department of Statistics, University of Munich.
** Munich School of Management, University of Munich.

August 2003

Abstract

Most econometric analyses of patent data rely on regression methods using a parametric form of the predictor for modeling the dependence of the response given certain covariates. These methods often lack the capability of identifying non-linear relationships between dependent and independent variables. We present an approach based on a generalized additive model in order to avoid these shortcomings. Our method is fully Bayesian and makes use of Markov Chain Monte Carlo (MCMC) simulation techniques for estimation purposes.

Using this methodology we reanalyze the determinants of patent oppositions in Europe for biotechnology/pharmaceutical and semiconductor/computer software patents. Our results largely confirm the findings of a previous parametric analysis of the same data provided by Graham, Hall, Harhoff & Mowery (2002). However, our model specification clearly verifies considerable non-linearities in the effect of various metrical covariates on the probability of an opposition. Furthermore, our semiparametric approach shows that the categorizations of these covariates made by Graham et al. (2002) cannot capture those non–linearities and, from a statistical point of view, appear to somehow ad hoc.

Keywords: Markov Chain Monte Carlo, Bayesian semiparametric binary regression, Latent utility models, Bayesian P–splines, Patent opposition.

1 Introduction

The analysis of patent data has a long tradition in economic research. The availability of large scale data sets drove researchers from different economic disciplines to address a broad range of questions by exploiting patent data in empirical analyses.

Traditional empirical work uses patent statistics as economic indicators linking patent data to some notion of inventiveness. This literature mainly deals with questions about the sources of economic growth, the rate of technical change, knowledge spill-overs or the competitive position of different firms and countries. Griliches (1990) provides a broad survey on patent statistics as economic indicators.

In a more specialized field, progress has been made in the difficult assessment of monetary value to patent rights driven by the analysis of citation and renewal data. Among the latest contributions to this field are Harhoff, Scherer & Volpel (1999) and Lanjouw, Pakes & Putnam (1998).

Most recently, empirical work focuses on the patent system itself. Especially the analysis of opposition and reexamination as well as litigation procedures attracted the interest of numerous researches. Harhoff & Reitzig (2002) and Graham et al. (2002) or Cockburn, Kortum et al. (2002) consider those legal actions as mechanisms to ensure a certain level of quality of issued patents while Lanjouw & Schankermann (2001) as well as Somaya (2003) interpret legal activities as good indicators for competition and conflict within different industries.

Existing empirical literature on patent data generally employs regression methods using a parametric form of the predictor for modeling the dependence of the response given certain covariates. Lanjouw & Lerner (1998) provide a comprehensive survey on recent empirical work. In this paper, we apply a semiparametric approach described in Fahrmeir & Lang (2001b) and Lang & Brezger (2003) to analyze the determinants and the effects of patent oppositions in Europe. Within a Bayesian framework we apply Markov Chain Monte Carlo methods (MCMC) for estimation purposes. In order to characterize the benefits from applying semiparametric models we compare our specification to the results of a simple parametric probit model employed by Graham et al. (2002) using their dataset on EPO patents from the biotechnology and pharmaceutical sector as well as from the semiconductor and computer software area.

Our results reveal some significant non-linearities in the effect of various covariates and show that the model specification of Graham et al. (2002) is not able to capture these non-linear effects correctly. Especially non-linearities in the effect of the number of states in which an invention seeks patent protection and in the effect of the number of a patent's forward citations leads to different results. Additionally, it turns out, that our semiparametric approach is superior to the parametric approach in terms of

the deviance information criterion (DIC) introduced by Spiegelhalter, Best, Carlin & van der Linde (2002), which can be used as a tool for model comparison in complex hierarchical Bayesian models and can be regarded as a Bayesian analogue to the Akaike information criterion.

Most of the methodology presented in this paper is implemented in *BayesX*, a software package for Bayesian generalized additive regression based on MCMC techniques. The program is available free of charge at

`http://www.stat.uni-muenchen.de/~lang`

The rest of the paper is structured as follows: Section 2 gives a brief review of the institutional background of patent opposition and litigation at the European Patent Office and summarizes previous findings from empirical studies of opposition/litigation activities. In Section 3 we discuss the Bayesian semiparametric regression framework and the MCMC simulation techniques which we use to analyze the data. Section 4 presents results from our semiparametric approach for modeling the probability of an opposition and compares them to the results of Graham et al. (2002). This section also includes a formal model comparison in terms of DIC. The paper closes with a short conclusion and some directions for further applications of the Bayesian semiparametric regression framework to the analysis of patent data.

2 The opposition mechanism of the European Patent Office

2.1 Institutional Background

From an economic point of view, the major purpose of a patent system is to spur innovation by providing the right incentives for innovative activity. Obtaining patent protection for an invention is equivalent to obtaining a temporary right to exclude others from using it. This allows the patent owner to benefit from the returns of his innovation while competitors are prohibited to copy the protected invention. In exchange for this temporary exclusion right the technical details of the underlying invention are made available to the public in the patent role. After the lapse of a patent any third party is allowed to copy and to commercially use the previously protected invention. Since welfare losses might be associated with the grant of patent protection, not every invention is suitable for patent protection. Only inventions which satisfy stringent patentability criteria can be protected by patents. A more detailed economic analysis of the economics of patent systems is given in Kaufer (1989).

In Europe, inventions which are seeking patent protection are examined (1) for their novelty, (2) their commercial applicability, (3) whether they mark an inventive step and (4) whether they are not excluded from

patentability for other reasons (European Patent Convention, 1973, Art. 52). Only inventions which satisfy these criteria can be protected by a European Patent. Patent applications at the European Patent Office (EPO) can be seen as a centralized process which leads to a bundle of individual patents in a subset of the 31 member and associated states of the EPC. Once a European patent is granted (and its validity is not challenged) it becomes a bundle of national patents in those states, which where specified in the application (European Patent Convention, 1973, Art. 3, 66, 79). According to the annual reports of the EPO about 65 % to 70 % of the applications at the EPO are granted.

Even if the examination process of the patentability of an invention is carried out by the patent examiner with the highest degree of diligence possible, it might lead to erroneous grant decisions. In order to correct such mistakes (and the associated welfare losses) most patent systems contain some post-grant mechanisms, which allows third parties to challenge the validity of granted patents. In general, patents can be challenged either within the patent office or before litigation courts. However, the possibilities of disputing a patent's validity differ considerably between patent systems.

Considering the EPO, any third party can oppose a patent by filing and substantiating an opposition within nine months after the grant decision, which is the case for about 8 to 10 % of all granted patents. An opposition can be substantiated by presenting evidence that one or more of the patentability criteria isn't satisfied by the protected invention. The opposition leads to one of three possible outcomes: the opposition may be rejected, the patent may be upheld with amendments or it may be revoked (European Patent Convention, 1973, Art. 101, 102). Once the nine months opposition period has lapsed, the validity of a patent can only be challenged in court. However, this may become a tedious and costly endeavor, since single suits have to be filed in each of the designated countries under the respective legal rules. A more detailed description of the possible legal procedures is given in Graham et al. (2002).

In the US there is no procedure comparable to the opposition mechanism of the EPO. The only possibility to challenge a patent's validity at the US Patent and Trademark Office (USPTO) is requesting a reexamination of the grounds upon which a patent was granted. Filing a reexamination requires the presentation of a previously undisclosed and relevant piece of prior art to the patent office. Any reexamination is proceeded if and only if it raises "a substantial new question of patentability" (United States Code Title 35, 2002, § 303 (a)) in the opinion of the examiner assigned by the USPTO. The patent office is required to make a determination of the validity of the patent if the reeaxmination goes forward (United States Code Title 35, 2002, § 307). During the procedure the patent-owner remains in contact with the USPTO and can offer amendments or new claims, while the role of possibly involved third parties remains limited. Reexamination can lead to the cancellation

of either some or even all or to the confirmation of either some or all of the claims specified in patent.

The validity of US patents can also be challenged in federal courts. Despite the high costs associated with going to court especially in the US, litigations are the dominant way of challenging a patent's validity in the US. Merges (1999) and Somaya (2003) find that only about 0.3 % of all issued US patents trigger reexamination while the rate of patents ending in courts is about 2 %.

A number of recent contributions observing a rapid growth in the number of so-called low-quality patents (patents that have been granted erroneously or with misspecified claims) in the US raised the interest in the efficiency of different systems of challenging a patent's validity. In this context, the reexamination system of the USPTO is widely seen as a rather inefficient mechanism for the correction of low-quality patents since it is too slow and too friendly to the patent-owner. According to Merges (1999) reexamination leads to a revocation of the challenged patent in only 12 % of the filed cases (the according revocation rate at the EPO is approximately 33 %). As a consequence it is rarely used. Some observers claim that the US system as a whole is an inefficient mechanism to correct for potential shortcomings in the examination of patent applications at the USPTO, since litigation in court as the second way of challenging a patent's validity also has drawbacks - it is a very costly and uncertain endeavor for the involved parties. In contrast, the EPO opposition system is seen as a 'role model'. Hall, Graham, Harhoff & Mowery (2003) as well as Levin & Levin (2002) argue that the adoption of an opposition system resembling the European system in the US could bring substantial welfare gains.

2.2 Empirical analysis of patent opposition

The current interest in the post-grant patent validity challenge came along with numerous empirical studies of the available mechanisms. The existing work mainly addresses incidence and outcomes of such procedures. Due to the infrequent use of the reexamination procedure at the USPTO, studies of challenging mechanisms for granted patents within patent offices focuses on the EPO opposition system. Among the most recent papers on this subject are Harhoff & Reitzig (2002) and Graham et al. (2002). Considering studies of litigation in courts, the contrary is true: Since European data is virtually not available, existing literature focuses on patent litigation in US federal courts as Lanjouw & Schankermann (2001) and Somaya (2003) did. A survey of the litigation literature can be found in Lanjouw & Lerner (1998).

The common methodology used in these papers is to model the probability of the occurrence of the discrete event 'opposition/litigation or not' dependent on a variety of patent indicators in order to analyze, which are the patents who are challenged more frequently than others. Among the

most prominent indicators is the number of citations made in the patent application (*backward citations*), the number of citations received by younger patents (*forward citations*), the number of claims stated in the patent (*claims*) and the number of states in which an innovation seeks patent protection (*designated states*). Additionally, measures of patent breadth as well as information on the filing strategy are usually included. Most of the indicators have been extensively discussed with respect to their theoretical and empirical validity in the literature on patent valuation. Interested readers are kindly addressed to the relevant sources for a detailed discussion of the current knowledge on patent indicators like Hall, Jaffe & Trajtenberg (2001). A detailed discussion on how to use patent statistics in order to build indicators for empirical analysis can be found in the Patent Manual of the OECD (1994). In the following, we briefly summarize the key findings for the influence of patent indicators on the incidence of a patent opposition giving also a short description of their economic interpretation. We limit ourselves to a description of the metrical indicators which are of primary interest in our analysis.

Citations: An inventor must cite all related prior patents and also non-patent literature within the patent application. During the examination process, the patent examiner is responsible for ensuring that all appropriate literature has been cited in the application, providing the right incentives that all relevant previous patents are cited in the application. It is generally assumed that *backward citations* (citations made in the application) operationalize existing market potential while *forward citations* (citations received by younger patents) are seen as a good indicator of a patent's social and monetary value. A detailed discussion on the economic interpretation of patent citations is found in Trajtenberg (1990). Econometric studies consistently find a significant positive influence of forward citations on the probability of the occurrence of opposition or litigation cases. Most recent studies comprise Lanjouw & Schankermann (2001), Harhoff & Reitzig (2002) and Graham et al. (2002). Harhoff & Reitzig (2002) argue, that given the cost of filing an opposition or litigation suit patents with higher economic value are more likely to be litigated than patents with a lower value.

Patent Claims: A patent comprises a set of *claims* that marks the boundaries of the patent. The principal claims state essential features of the underlying invention while subordinate claims usually describe detailed features of the innovation. Lanjouw & Schankermann (1999) interpret the number of claims as one measure of a patent's breadth and they find that this measure is highly correlated with the value of a patent. Additionally, Harhoff & Reitzig (2002) and Lanjouw & Schankermann (2001) find that the number of

claims in a patent significantly rises the probability of an opposition respective litigation. Again the rationale is that the number of claims is correlated with the value of patents and that valuable patents are more likely to be litigated.

Designated States: The *number of designated states* (or the *'family size'* of a patent) is equivalent to the number of jurisdictions in which patent protection is sought. The number of designated states can be used as a measure for the territorial size of a patent. Lanjouw et al. (1998) find a strong correlation between the number of designated states and the life span of a patent. They argue that the number of states is positively correlated with the value of patents (which is confirmed in Harhoff et al. (1999)) and more valuable patents are more likely to be prolongated, since prolongation is costly to the patent holder.

A variety of other indicators has been used as covariates in the analysis of patent litigation. Among those are patent breadth, ownership variables (mainly whether the owner of a patent is an individual, a corporation or a university) and indicators referring to the filing strategy of the patent applicant (indicators whether an accelerated search or examination of the application was requested by the applicant and whether a PCT application has been filed). However, empirical evidence of the validity of these indicators varies among different studies.

Note that the use of classical parametric regression methods is a common feature of the largest part of the empirical literature on patent opposition/litigation. However, the recent advances in computing power and the structure of the available data makes it possible to estimate more flexible models. In the following, we present a semiparametric model for the probability of an opposition. The unknown parameters and functions of the model are estimated in a Bayesian framework using Markov Chain Monte Carlo simulation techniques. Furthermore, we show that our approach has major advantages over pure parametric specifications.

3 Bayesian semiparametric binary regression

3.1 Structural assumptions

Consider regression situations, where observations (y_i, z_i), $i = 1, \ldots, n$, on a binary response y and covariates z are given, which can be divided into metrical covariates $x_1, \ldots, x_p$ and categorical covariates $w_1, \ldots, w_q$. The most widely used models for binary data are logit or probit models. Given the covariates the responses y_i are binomially distributed, i.e. $y_i|z_i \sim B(1, \pi_i)$ with the probability of success $\pi_i = P(y_i = 1|z_i) = E(y_i|z_i)$ being modeled

as

$$\pi_i = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)}$$

for logit models or

$$\pi_i = \Phi(\eta_i)$$

for probit models. Here, η_i is the predictor that models the influence of the covariates on the probability π_i.

An alternative way of obtaining a probit model, which is very useful for Bayesian inference, is to express binary regression models in terms of latent utilities, see e.g. Fahrmeir & Tutz (2001) or Fahrmeir & Lang (2001b). Introducing the metric latent utilities

$$U_i = \eta_i + \epsilon_i$$

with i.i.d. errors ϵ_i, we define $y_i = 1$ if $U_i > 0$ and $y_i = 0$ if $U_i < 0$. Then, the assumption $\epsilon_i \sim N(0, 1)$ yields the well known probit model.

Concerning the form of the predictor and the type of the influence of metrical covariates $x_1, \ldots, x_p$ the following three approaches will be distinguished for the rest of the paper

Setting M_1: In the simplest approach the effects of the metrical covariates $x_1, \ldots, x_p$ are incorporated into the model by additive linear terms $x_1'\beta_1, \ldots, x_p'\beta_p$. The predictor can then be written by

$$\eta_1 = \sum_{j=1}^{p} x_j'\beta_j + w'\gamma$$

with the unknown regression parameters given by $\theta = (\beta_1, \ldots, \beta_p, \gamma)$.

Setting M_2: In many practical situations, as in our application on patent opposition data, the assumption of linear effects of the metrical covariates on the predictor is too restrictive. A simple and widely used way to allow for non–linearities in the effects of metrical covariates is to categorize some or all $x_1, \ldots, x_p$ and then construct a set of r_j dummy variables $\tilde{x}_j, j = 1, \ldots, p$. The linear terms $x_j'\beta_j$ are then replaced by $\tilde{x}_j'\tilde{\beta}_j$ with $\tilde{\beta}_j = (\tilde{\beta}_{j1}, \ldots, \tilde{\beta}_{jr_j})'$ and the predictor can be defined by

$$\eta_2 = \sum_{j=1}^{p} \tilde{x}_j'\tilde{\beta}_j + w'\gamma$$

with the unknown regression parameters $\theta = (\tilde{\beta}_1, \ldots, \tilde{\beta}_p, \gamma)$. Note that in this setting the number r_j and location of the intervals defining the components of the dummy vector $\tilde{x}_j$ has a crucial influence on the degree and shape of non–linearity in the estimated effect. In general, increasing r_j leads to more flexible regression effects $\tilde{\beta}_j$ but also to an inflation in the number of effective parameters which have to be estimated and interpreted.

Setting M_3: An alternative, more flexible and data-driven method for modeling non–linear effects of metrical covariates is to incorporate them additively into the predictor by using smooth regression functions $f_j(x_j)$ instead of linear terms $x_j\beta_j$ or $\tilde{x}_j'\tilde{\beta}_j$. This leads to a semiparametric additive predictor of the form

$$\eta_3 = \sum_{j=1}^{p} f_j(x_j) + w'\gamma$$

where we assume possibly nonlinear effects $f_1, \ldots, f_p$ for the metrical covariates. The unknown parameters are given by $\theta = (f_1(x_1), \ldots, f_p(x_p), \gamma)$ with $f_j(x_j)$ representing a vector of function evaluations. Compared to M_2 the semiparametric approach allows for the modeling of very complex, non–linear regression functions without suffering from the parameter inflation problem if a very flexible effect has to be estimated. Furthermore, the degree of non–linearity does not have to be predefined by choosing the number and location of the categories in the construction of the set of dummy variables $\tilde{x}_j$, but can be estimated jointly with the unknown regression parameters depending only on the observed data.

Note that M_2 can be regarded as a special case of M_3 by choosing a step function defined on given categorization intervals as the regression function and that we omitted the intercept term γ_0 in the predictors notationally, which is tacitly assumed to be included in the parametric part $w'\gamma$.

To demonstrate the differences between our three approaches, we want to present some preliminary results from the analysis of EPO patent opposition data discussed in more detail in Section 4. For our example, the probability of the occurence of an opposition is modeled only depending on the number of designated states, a metrical covariate. Figure 1 (a) shows the empirical rate of opposition plotted against the number of designated states and indicates that the probability for an opposition is higher for more designated states with a small drop for 12 to 14 states.

To model this probability, in M_1 the effect of the number of designated states is incorporated into the predictor by a simple linear term. Following the example of Graham et al. (2002) the set of dummy variables in M_2 is constructed by categorizing the number of states into the three categories "less than 6" (reference category), "between 6 and 10", and "more than 10". For M_3 a nonparametric regression function with a P–spline approach described in more detail in Section 3.2 is used. The estimation of the unknown parameters in all three cases is fully Bayesian and will be explained in Section 3.2.

Figure 1 (b) shows the estimated probabilities for M_1, M_2, M_3 and reveals that only the semiparametric approach M_3 is capable of detecting the drop in opposition rate for 12 to 14 designated states. Furthermore it is obvious that both M_1 and M_2 are not able to capture the underlying depen-

dence structure between opposition probability and number of designated states as accurately as M_3 does.

3.2 Bayesian inference via Markov Chain Monte Carlo

Prior assumptions

In a Bayesian approach unknown functions $f_1, \ldots, f_p$ and parameters $\beta = (\beta_1, \ldots, \beta_p)$, $\tilde{\beta} = (\tilde{\beta}_1, \ldots, \tilde{\beta}_p)$, γ of fixed effects are considered as random variables and have to be supplemented by appropriate prior distributions.

In the absence of any prior knowledge a typical assumption for the parameters of the fixed effects is to use independent diffuse priors, i.e. $p(\beta) \propto const$, $p(\tilde{\beta}) \propto const$ and $p(\gamma) \propto const$.

For the unknown regression functions f_j, we will use a P–splines approach originally introduced by Eilers & Marx (1996) and formulated in a Bayesian setting by Lang & Brezger (2003). In a P–splines approach it is assumed that the unknown functions f_j can be approximated by linear combinations

$$f_j(x_j) = \sum_{k=1}^{m_j} \delta_{jk} B_{jk}(x_j)$$

of $m_j = l_j + r_j$ linearly independent B–spline basis functions B_{jr} of degree l_j defined on a set of r_j equally spaced knots $x_{j,min} = \xi_{j0} < \ldots < \xi_{jr_j} = x_{j,max}$. The basis functions can be regarded to have compact local support in the sense that they are nonzero only on a domain spanned by the $l_j + 2$ knots, whereas the B–spline coefficients $\delta_j = (\delta_{j1}, \ldots, \delta_{jm_j})'$ act as weights assigned to each single basis function.

To ensure both enough flexibility and sufficient smoothness of the fitted curves, Eilers & Marx (1996) proposed to use a relatively large number of knots but, in order to prevent overfitting, to penalize adjacent B–spline coefficients with differences of order d. In their frequentist approach this leads to penalized likelihood estimation with roughness penalties, where the penalized likelihood is maximized with respect to the unknown regression parameters and the trade off between flexibility and smoothness is controlled by additional smoothing parameters λ_j. In general, large values of λ_j give large weight to the roughness penalty term thereby enforcing smooth functions, while for small λ_j, the functions tend to be closer to the data.

In a Bayesian setting, the difference penalties are replaced by their stochastic analogues, i.e. random walks of order d. For simplicity, we will restrict to $d = 2$, which corresponds to a second order random walk

$$\delta_{jk} = 2\delta_{j,k-1} - \delta_{j,k-2} + u_{jk}$$

for adjacent B–splines coefficients δ_{jk} with Gaussian errors $u_{jk} \sim N(0, \tau_j^2)$ and diffuse priors $p(\delta_{j1})$ and $p(\delta_{j2}) \propto const$ for initial values. Note, that this prior may be equivalently defined in a symmetric form by specifying

the conditional distributions of a particular parameter δ_{jk} given its left and right neighbors. Then, for $d = 2$, the conditional means can be interpreted as locally quadratic fits at the knot positions ξ_{jk}, see e.g. Besag, Green, Higdon & Mengerson (1995). The amount of smoothness is controlled by the additional variances τ_j^2, which are related to the smoothness parameters λ_j in a frequentist approach by $\lambda_j = (\tau_j^2)^{-1}$. Thus, larger (smaller) values for the variances lead to rougher (smoother) estimates for the regression function.

The joint prior of the B–splines coefficients δ_j is Gaussian and can easily be computed as

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

with a penalty matrix $K_j = D'D$, where D is a second order difference matrix. For second order random walks, for example, K_j is given by

$$K_j = \begin{pmatrix}
1 & -2 & 1 & & & & & & \\
-2 & 5 & -4 & 1 & & & & & \\
1 & -4 & 6 & -4 & 1 & & & & \\
 & 1 & -4 & 6 & -4 & 1 & & & \\
 & & \ddots & \ddots & \ddots & \ddots & \ddots & & \\
 & & & 1 & -4 & 6 & -4 & 1 & \\
 & & & & 1 & -4 & 6 & -4 & 1 \\
 & & & & & 1 & -4 & 5 & -2 \\
 & & & & & & 1 & -2 & 1
\end{pmatrix}$$

with zero elements outside the second off–diagonals.

For a fully Bayesian analysis, variance or smoothness parameters τ_j^2 are also considered to be unknown and estimated simultaneously with the unknown regression parameters. Therefore, hyperpriors are assigned to them in a second stage of the hierarchy by assuming highly dispersed inverse gamma distributions $\tau_j^2 \sim IG(a_j, b_j)$ with known hyperparameters a_j and b_j. A common choice for the hyperparameters is $a_j = 1$ and $b_j = 0.0005$ leading to an almost diffuse prior for τ_j^2. Results are rather insensitive to the choice of a_j and b_j for moderate to large datasets, but, for small datasets, a sensitivity analysis is useful. Note, that these prior assumptions for the smoothness parameters are a major advantage over a classical frequentist approach, where smoothness parameters usually have to be specified by hand or a complex grid search algorithm has to be performed.

Posterior analysis

As it would be beyond the scope of this paper to present a detailed introduction into Bayesian inference using MCMC in semiparametric regression models for binary response, we will focus only on some key results given

in Fahrmeir & Lang (2001b) and Lang & Brezger (2003). For a thorough treatment of MCMC in general refer, for example, to Green (1999) or Gilks, Richardson & Spiegelhalter (1996).

Bayesian inference is based on the posterior and is carried out using recent MCMC simulation techniques. Let θ denote the vector of all unknown parameters in the model. Then, under usual conditional independence assumptions, the posteriors augmented by the latent variables for the three approaches described in Section 3.1 are given by

$$\begin{aligned}
M_1: \quad p(\theta|Y) &\propto p(Y|U) \cdot p(U|\eta) \cdot p(\beta) \cdot p(\gamma) \\
M_2: \quad p(\theta|Y) &\propto p(Y|U) \cdot p(U|\eta) \cdot p(\tilde{\beta}) \cdot p(\gamma) \\
M_3: \quad p(\theta|Y) &\propto p(Y|U) \cdot p(U|\eta) \cdot \prod_{j=1}^{p} \{p(\delta_j|\tau_j^2)p(\tau_j^2)\} \cdot p(\gamma)
\end{aligned}$$

Because the direct maximization of all three posterior distributions is not possible, MCMC methods have to be applied in order to be able to estimate the unknown parameters β, $\tilde{\beta}$, γ, δ_j and τ_j^2, which make use of the full conditionals, i.e. the distribution of a certain parameter block given all the other parameters.

The full conditionals for the fixed effects parameters β, $\tilde{\beta}$ and γ as well as for the parameter vectors $\delta_1, \ldots, \delta_p$ are multivariate Gaussian. For the variance components τ_j^2 the full conditionals are inverse gamma distributions. Finally, it can be shown that the full conditionals of the latent variables U are truncated normals, subject to the constrains $U_t > 0$ if $y_t = 1$ and $U_t < 0$ if $y_t = 0$.

Thus, a Gibbs sampler originally introduced by Geman & Geman (1984) can be used for MCMC simulation, drawing successively from the full conditionals for the latent variables U, for the fixed effects parameters β, $\tilde{\beta}$ and γ, for the B–splines coefficients δ_j and for the variances τ_j^2. Running this Gibbs sampler yields random samples from the marginal distributions of the regression parameters β, $\tilde{\beta}$, γ, δ_j and τ_j^2, from which Bayesian point estimates like posterior means or posterior medians can be calculated. Additionally, in order to assess the significance of the estimates, it is possible to compute credible regions, an analogue to confidence intervals in a frequentist approach, by calculating suitable quantiles based on the obtained random samples.

4 Analysis of patent opposition at the EPO

In this section we reinvestigate a dataset of approximately 4800 patents from the biotechnology/pharmaceutical and semiconductor/computer software sectors granted by the EPO between 1980 and 1997, which has previously been analyzed by Graham et al. (2002). For reasons of brevity we skip descriptive details of the data–set, which are given in the original paper.

The aim is to model the probability that an opposition against a granted patent occurs yielding the binary response variable

$$\begin{aligned} y_i = 1 &\quad \Leftrightarrow \quad \text{Opposition} \\ y_i = 0 &\quad \Leftrightarrow \quad \text{No opposition} \end{aligned}$$

As our main focus was to show that a semiparametric regression approach does have clear benefits compared to a simple linear probit model, we used only the significant covariates found by Graham et al. (2002), which are summarized in Table 1.

As a first step for modeling the probability of an opposition given the covariates, we use a simple linear model M_1 with the predictor

$$\eta_1 = \beta_0 + x_1'\beta_1 + x_2'\beta_2 + x_3'\beta_3 + x_4'\beta_4 + w'\gamma$$

where the influence of the metrical covariates is assumed to be linear. The estimation results for the unknown regression parameters in this setting are given in Table 3 (a) and show, that the probability for an opposition decreases over time, but increases with increasing number of EPO forward citations, number of EPO claims and number of designated states. These results are in line with previous findings described in Section 2.2. Concerning the effect of the binary covariates $w_1, \ldots, w_7$ it turns out, that the opposition probability is higher for patents from the biotech/pharmaceutical sector, for patents with a patentholder from Switzerland, Germany or Great Britain, for patents with an accelerated examination request and for patents with a PCT filing. Adversely, for patents with a US twin, for patents with a patentholder from the US and for patents with an accelerated search request, the probability for an opposition is reduced. The computed 95 % credible regions for the estimated parameters given in Table 3 (b) and (c) do not include zero, so all effects are significant on the 5 % error level. Finally, Table 3 (d), summarizes the marginal changes in probability for a unit change of the covariate/dummy if all other covariates are set to zero. For example, each additional designated state increases the opposition probability by approximately 1.8 % while for patents with a US twin this probability is lowered by approximately 8 % compared to a patent with no US twin.

Extending this fully linear model in order to incorporate possible non–linearities in the effects of the metrical covariates $x_1, \ldots, x_4$, we now compare M_1 to the approach M_2 used by Graham et al. (2002) with a set of dummy effects for categorized versions of the metrical covariates and to our semiparametric approach M_3, where smooth regression functions $f_1(x_1), \ldots, f_4(x_4)$ are used. The predictors can then be defined by

$$\begin{aligned} \eta_2 &= \beta_0 + \tilde{x}_1'\tilde{\beta}_1 + \tilde{x}_2'\tilde{\beta}_2 + \tilde{x}_3'\tilde{\beta}_3 + \tilde{x}_4'\tilde{\beta}_4 + w'\gamma \\ \eta_3 &= \beta_0 + f_1(x_1) + f_2(x_2) + f_3(x_3) + f_4(x_4) + w'\gamma \end{aligned}$$

with the dummy vectors $\tilde{x}_2, \ldots, \tilde{x}_4$ based on the categories given in Graham et al. (2002) and summarized in Table 2. Differing slightly from Graham et al. (2002), we only used 9 biannual categories for the grant year represented in M_3 by $\tilde{x}_1$.

Figure 2 displays the estimated effects of the metrical covariates for both M_2 and M_3. Note that the effects have been centered appropriately to ensure identifiability and comparability. Roughly speaking, the results for the metrical covariates are similar to the ones obtained from M_1, but it is obvious that especially the effects for the number of designated states depicted in Figure 2 (b) and the number of EPO forward citations depicted in Figure 2 (c) are clearly non–linear and that the dummy effects obtained from M_2 are very raw approximations of the true underlying dependency structure represented by the smooth effects in M_3. Additionally, Figure 2 (d) shows, that especially for the number of a patent's EPO claims the categorization used by Graham et al. (2002) is not chosen very well in putting all patents with more than 15 EPO claims into one category with a constant effect. In fact, for patents with more than about 30 EPO claims the estimated smooth effect $f_4(x_4)$ indicates that the probability for an opposition increases dramatically and is much higher than indicated by the corresponding dummy effect. The significance of the smooth effects in M_3 is supported by the pointwise 95 % credible regions also depicted in Figure 2, which are clearly different from zero for most values of the corresponding covariate. Concerning the results for the binary covariates $w_1, \ldots, w_7$ we will omit a detailed discussion for both M_2 and M_3 as they are similar to the results obtained from the fully linear model M_1 presented in Table 3.

To give a more formal rationale for the benefits in using our semiparametric approach, we compared the three approaches M_1, M_2, M_3 in terms of the deviance information criterion (DIC) introduced by Spiegelhalter et al. (2002). The DIC is a Bayesian analogue to the Akaike information criterion penalizing the fit of a model measured by the deviance with the complexity of a model represented by the effective number of model parameters. The results are given in Table 4 and show, that the DIC is clearly minimized by the semiparametric approach M_3 and that the approach M_2 used by Graham et al. (2002) is in fact even worse than the simple fully linear probit model M_2.

5 Conclusions and further work

In this paper, we have used a Bayesian semiparametric regression approach to model the probability of an opposition for EPO patents from the biotechnology/pharmaceutical and semiconductor/computer software sectors. The opposition probability turned out to increase with increasing number of designated states, number of EPO patent claims and number of EPO forward

citations, but, unlike previous researchers, we could show that this increase was clearly non–linear by incorporating the effects of these metrical covariates in form of smooth regression functions instead of simple linear terms. Due to the hierarchical structure of our Bayesian approach, the smoothness of the estimated functions is totally data–driven and estimated jointly with the unknown regression parameters. A formal model comparison based on the deviance information criterion (DIC) supported the benefits of our approach compared to a fully parametric model used by Graham et al. (2002).

One focus for future research could be a segmentation routine detecting similarities in patent/opposition characteristics independent of prespecified technology or geographical classifications based on an extension of Bayesian additive mixed models. Additionally, the application of Bayesian semiparametric models for multicategorical response and for survival analysis might be useful in the analysis of the outcome as well as the duration of the opposition procedure. For an introduction into the named model classes refer to Fahrmeir & Lang (2001a), Fahrmeir & Lang (2001b) and Hennerfeind, Brezger & Fahrmeir (2003).

Acknowledgement

This research was supported by the German National Science Foundation (DFG), Sonderforschungsbereich 386 "Statistical Analysis of Discrete Structures". We would like to thank Ludwig Fahrmeir and Dietmar Harhoff for helpful discussions.

References

Besag, J., Green, P., Higdon, D. & Mengerson, K. (1995). Bayesian computation and stochastic systems (with discussion), *Statistical Science* **10**: 3–66.

Cockburn, L., Kortum, S. et al. (2002). Are all patent examiners equal? The impact of characteristics on patent statistics and litigation outcomes, *Working Paper 8980*, NBER.

Eilers, P. & Marx, B. (1996). Flexible smoothing using B–splines and penalized likelihood (with comments and rejoinder), *Statistical Science* **11**: 89–121.

European Patent Convention (1973). *URL:* http://www.european-patent-office.org/legal/epc/e/ma1.html#CVN.

Fahrmeir, L. & Lang, S. (2001a). Bayesian inference for generalized additive mixed models based on markov random field priors, *Journal of the Royal Statistical Society C (Appl. Stat.)* **50**: 201–220.

Fahrmeir, L. & Lang, S. (2001b). Bayesian semiparametric regression analysis of multicategorical time–space data, *Annals of the Institute of Statistical Mathematics* **53**: 10–20.

Fahrmeir, L. & Tutz, G. (2001). *Multivariate Statistical Modelling based on Generalized Linear Models*, 2nd edn, Springer–Verlag, New York.

Geman, S. & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, *IBEE Transactions on Pattern Analysis and Machine Intelligence* **6**: 721–741.

Gilks, W. R., Richardson, S. & Spiegelhalter, D. J. (eds) (1996). *Markov Chain Monte Carlo in Practice*, Chapman and Hall, London.

Graham, S., Hall, B., Harhoff, D. & Mowery, D. (2002). Post-issue patent "quality control": A comparative study of US patent reexaminations and European patent oppositions, *Working Paper 8807*, NBER.

Green, P. J. (1999). A primer on Markov Chain Monte Carlo, *in* O. E. Barndorff-Nielsen, D. R. Cox & C. Klüppelberg (eds), *Complex Stochastic Systems*, Chapman and Hall, London, pp. 1–62.

Griliches, Z. (1990). Patent statistics as economic indicators, *Journal of Economic Literature* **28**(4): 1661–1707.

Hall, B., Graham, S., Harhoff, D. & Mowery, D. (2003). Prospects for improving U.S. patent quality, *Working Paper 9341*, NBER.

Hall, B., Jaffe, A. B. & Trajtenberg, M. (2001). The NBER patent citations data file: Lessons, insights and methodological tools, *Working Paper 8498*, NBER.

Harhoff, D. & Reitzig, M. (2002). Determinants of opposition against EPO patent grants - The case of biotechnology and pharmaceuticals, *Discussion Paper 3645*, Centre for Economic Policy Research.

Harhoff, D., Scherer, F. & Volpel, K. (1999). Citations, family size, opposition and the value of patents rights, *Discussion paper*, University of Munich.

Hennerfeind, A., Brezger, A. & Fahrmeir, L. (2003). Geoadditive survival models, *Discussion Paper 333*, SFB 386, University of Munich.

Kaufer, E. (1989). *The Economics of the Patent System*, Harwood Academic Publishers GmbH, New York.

Lang, S. & Brezger, A. (2003). Generalized structured additive regression based on Bayesian P–splines, *Discussion Paper 321*, SFB 386, University of Munich.

Lanjouw, J. O. & Lerner, J. (1998). The enforcement of intellectual property rights: A survey of the empirical literature, *Annales d'Economie et de Statistiques* **49/50**: 223–246.

Lanjouw, J. O., Pakes, A. & Putnam, J. (1998). How to count patents and value intellectual property: Uses of patent renewal and application data, *Journal of Industrial Economics* **46**(4): 405–433.

Lanjouw, J. O. & Schankermann, M. (1999). The quality of ideas: Measuring innovation with multiple indicators, *Working Paper 7345*, NBER.

Lanjouw, J. O. & Schankermann, M. (2001). Characteristics of patent litigation: a window on competition, *RAND Journal of Economics* **32**(1): 129–151.

Levin, J. & Levin, R. (2002). Patent oppositions, *Discussion Paper 01-29*, Stanford Institute for Economic Policy Research.

Merges, R. P. (1999). As many as six impossible patents before breakfast: Property rights for business concepts and patent system reform, *Berkeley Technology Law Journal* **14**: 577–615.

OECD (1994). Using patent data as science and technology indicators - Patent manual 1994. Paris.

Somaya, D. (2003). Strategic determinants of decisions not to settle patent litigation, *Strategic Management Journal* **24**: 17–38.

Spiegelhalter, D., Best, N., Carlin, B. & van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion), *Journal of the Royal Statistical Society B* **64**(4): 583–639.

Trajtenberg, M. (1990). A penny for your quotes: Patent citations and the value of innovations, *RAND Journal of Economics* **21**(1): 172–187.

United States Code Title 35 (2002). Consolidated patent laws. *URL:* http://www.uspto.gov/web/offices/pac/mpep/consolidated_laws.pdf.

Address:
Alexander Jerak, Department of Statistics, University of Munich, Ludwigstr. 33, 80539 Munich, Germany. email: jerak@stat.uni-muenchen.de
Stefan Wagner, Department of Business Administration, INNO-tec, University of Munich, Kaulbachstr. 45, 80539 Munich, Germany. email: swagner@bwl.uni-muenchen.de

Metrical covariates	
x_1	Grant year
x_2	Number of EPO forward citations
x_3	Number of designated states
x_4	Number of EPO claims

Binary covariates (1 = Yes / 0 = No)	
w_1	Patent from biotech/pharma sector
w_2	US twin exists
w_3	Patentholder from US
w_4	Patentholder from Switzerland, Germany or Great Britain
w_5	Accelerated exam requested
w_6	Accelerated search requested
w_7	PCT filing

Table 1: EPO patent opposition. Summary of covariates.

Grant year (x_1)
9 categories $\{1980/1981\}$, $\{1982/1983\}$,..., $\{1996/1997\}$ with reference category $\{1980/1981\}$

Number of EPO forward citations (x_2)
5 categories $\{0\}$, $\{1\}$, $\{2-5\}$, $\{6-10\}$, $\{>10\}$ with reference category $\{0\}$

Number of designated states (x_3)
3 categories $\{<6\}$, $\{6-10\}$, $\{>10\}$ with reference category $\{<6\}$

Number of EPO claims (x_4)
5 categories $\{<6\}$, $\{6-9\}$, $\{10\}$, $\{11-15\}$, $\{>15\}$ with reference category $\{<6\}$

Table 2: EPO patent opposition. Summary of categories for metrical covariates in M_2.

Covariate	(a) Posterior Mean	(b) 2.5%-Quant.	(c) 97.5%-Quant.	(d) $d\pi/dx$
Intercept	-0.4392	-0.6053	-0.2828	
x_1	-0.0479	-0.0586	-0.0381	-1.7 %
x_2	0.0915	0.0721	0.1105	$+3.4$ %
x_3	0.0492	0.0363	0.0613	$+1.8$ %
x_4	0.0133	0.0084	0.0180	$+0.5$ %
w_1	0.3696	0.2692	0.4697	$+14.2$ %
w_2	-0.2363	-0.3178	-0.1524	-8.0 %
w_3	-0.1418	-0.2305	-0.0535	-4.9 %
w_4	0.1760	0.0747	0.2759	$+6.6$ %
w_5	0.5992	0.3531	0.8472	$+23.3$ %
w_6	-0.3760	-0.7312	-0.0390	-11.9 %
w_7	0.2754	0.1708	0.3810	$+10.5$ %

Table 3: EPO patent opposition. Results for model M_1. (a) Posterior mean estimate of regression parameter. (b) Lower value of 95 % credible region. (c) Upper value of 95 % credible region. (d) Marginal change in probability for a unit change of the covariate/dummy.

	Deviance	pD	DIC
M_1	5680.63	11.88	5704.39
M_2	5671.49	25.41	5722.31
M_3	5629.36	30.89	5691.14

Table 4: EPO patent opposition. Comparison of deviance, effective number of parameters (pD) and DIC for models M_1, M_2 and M_3.

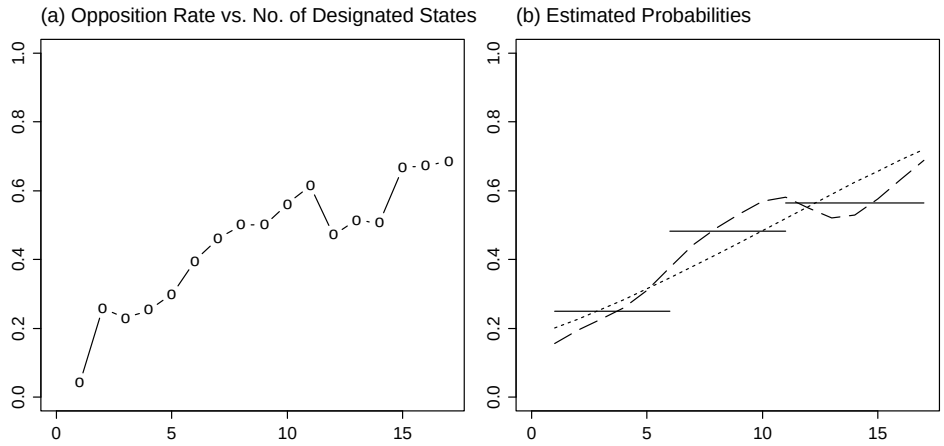

Figure 1: Comparison of considered model settings. (a) Rate of opposition versus number of designated states. (b) Probabilities of opposition estimated with M_1 ($\cdot \cdot \cdot$), M_2 (—), M_3 (- - -) and number of designated states as covariate.

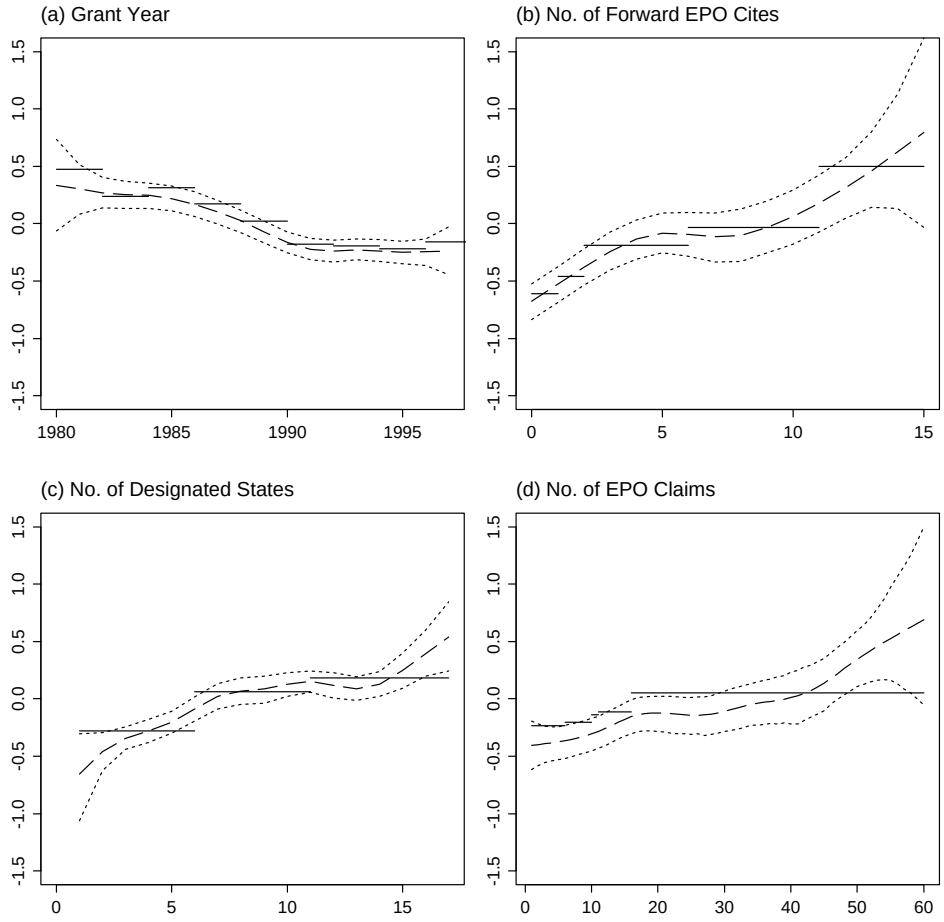

Figure 2: EPO patent opposition. Results for effect of (a) grant year, (b) number of forward EPO cites, (c) number of designated states, (d) number of EPO claims. Shown is for model M_2 the posterior mean (—) of the corresponding dummy effects, for model M_3 the posterior mean (- - -) of the corresponding regression function within 95 % credible regions ($\cdots$).