Discussion of 'Beyond Mean Regression' by Thomas Kneib

Peter J. Green[*]
University of Bristol and University of Technology, Sydney

February 10, 2013

Abstract

Methodology for regression beyond the mean has been a goal of researchers for many years. This discussion provides some additional context for the important ideas in the present paper, by recounting some of the historical background to the GAMLSS approach and pointing to the power and appeal of fully probabilistic regression analysis in the setting of Bayesian nonparametrics.

Some key words: Back-fitting, Bayesian nonparametrics, Dirichlet process, LMS method, partial splines, penalised likelihood.

1 Introduction

My discussion focusses on two themes, and in both, I give a personal view. One theme, at the Editor's suggestion, gives a little more of the historical background on some of the progenitors of the GAMLSS (Rigby and Stasinopoulos 2005) methodology that underpins a substantial part of Kneib's paper; this history intersects with a few contributions of mine in semi-parametric model fitting, penalised likelihood, etc. The second theme describes some recent approaches to 'beyond mean regression' inferential problems from the standpoint of Bayesian nonparametrics – the paradigm that I would probably adopt if I was working in the area now. Although still rather in their infancy, such methods promise to deliver a very comprehensive solution to the inferential problems raised in this paper, so I hope this provides a useful additional perspective.

2 Looking back

2.1 Partial splines and back-fitting

My first venture into semiparametric regression, although I didn't use that term at the time, was some joint work (Green *et al.* 1985) that aimed at a particular context – the analysis of agricultural

[*] School of Mathematics, University of Bristol, Bristol BS8 1TW, UK., and
School of Mathematical Sciences, University of Technology, Sydney, Australia.
Email: P.J.Green@bristol.ac.uk.

field trials. This is of course one of the earliest domains of application in statistical inference, one where many of the most fundamental ideas of our subject including the design of experiments, linear models, analysis of variance and covariance, randomisation, etc., were first explored. This very classical area was rather intensively re-visited in the 1980's by researchers who were interested in adjusting estimates of treatment or variety effects to allow for the effect of variations in 'fertility' across the physical extent of the field where the experiment was laid out. A generic formulation of the problem takes the form

$$Y = X\beta + g + \varepsilon \tag{2.1}$$

where $y = (y_i)_{i=1}^n$ is the vector of crop yields indexed by plot index i, X is the $n \times p$ design matrix for treatment, variety and other fixed effects, β the p-vector of corresponding parameters, $g = (g_i)_{i=1}^n$ the so-called fertility trend and ε a vector of errors. Methods of analysis of data in this set-up differ mainly according to the philosophy and structure of the model used for g. The standard methods of analysis exploited various complete and incomplete block designs, so that, effectively, spatial variations in fertility were adjusted for at the macro level by the choice of blocks, and at the micro level by relying on the randomisation in the layout. So-called 'neighbour' methods effectively model g in an explicit and more continuous manner, for example using ideas from geostatistics or spatial processes.

New approaches to these questions were promoted by researchers who did not necessarily draw exclusively on this classical tradition, but had other kinds of background, notably in spatial statistics. The seminal paper by Wilkinson *et al.* (1983) stimulated a discussion that included many novel ideas; my colleagues and I proposed a method we called 'least squares smoothing' which was subsequently presented in full detail in Green *et al.* (1985). In this approach, the decomposition (2.1) is fitted by minimising the penalised sum of squares

$$\sum_{i=1}^n (Y_i - (X\beta)_i - g_i)^2 + \lambda \sum (\Delta g)_j^2 \tag{2.2}$$

using (typically) the second difference operator Δ, within each contiguous line of plots in a one-dimensional spatial array. The tuning parameter $\lambda \in (0, \infty)$ determines the trade-off between fit and smoothness.

This work was conducted independently of the partial spline approach to semiparametric regression being explored by Wahba and others (Wahba 1990) at the same time. There, the model under consideration might be written

$$Y_i = (X\beta)_i + g(t_i) + \varepsilon_i, \quad i = 1, \dots, n$$

and the estimation criterion is to minimise

$$\sum_{i=1}^n (Y_i - (X\beta)_i - g(t_i))^2 + \lambda \int (g''(t))^2 dt; \tag{2.3}$$

the parallel is obvious. The partial spline model has found enormously wide use in applied statistical methodology, and has stimulated a rich and fruitful corpus of research leading to Wahba (1990) and beyond. See also Green and Silverman (1994), chapter 4. Although penalised least squares is almost universally used as the estimation criterion, this does generate an interesting and surprising puzzle in inference: in general the bias in estimating β in this way can be asymptotically larger than its standard error (see Rice (1986), and, for a possible solution, Speckman (1988)). It is curious that this fact has been widely ignored in study and use of semiparametric regression methods ever since.

Computing the penalised least squares estimates of β and g in either the discrete (2.2) or continuous (2.3) formulation of the partial spline model can be organised in several ways. Green *et al.* (1985) described how either β or g could be eliminated from the normal equations for the model, resulting in linear systems of dimension n or p respectively. But more commonly, models like this are fitted by alternately minimising the penalised sum of squares with respect to β and g, until convergence is obtained; this is ensured through a simple spectral analysis, e.g. Green and Silverman (1994), Theorem 4.2. This iterative method is called *back-fitting*, see also Breiman and Friedman (1985) and Buja *et al.* (1989), and is of course the key computational tool also in fitting more general additive models.

2.2 Penalised likelihood

If we interpret the error sum of squares in (2.2) or (2.3) as a negative scaled log-likelihood, then we can immediately see how to generalise the penalised least squares idea to non-normal models. This idea was explored in Green (1987), and treated more thoroughly in the context of generalised linear models in Green and Silverman (1994), chapter 5. This approach is called maximum penalised likelihood, and the objective is to maximise

$$\ell(\theta, \phi) - \tfrac{1}{2}\lambda \int g''(t)^2 dt$$

(we drop the discrete formulation from now on). Here ℓ is the log-likelihood for a generalised linear model with natural parameter θ and common dispersion parameter ϕ; θ is related to the linear predictor $\eta = X\beta + (g(t_i))_{i=1}^n$ through a link function G, with $G(b'(\theta_i)) = \eta_i$ as usual. We have rescaled λ to account for the dispersion parameter (the variance in the normal case). Exactly the same penalised likelihood approach can be taken not only for generalised linear models but much more broadly: it really applies whenever the unknowns in the likelihood are a vector of parameters β and a finite set of linear functionals of a smooth curve g; none of linearity, additivity or an exponential family distribution are truly essential.

Again, backfitting is the usual method of computing the maximum penalised likelihood estimates, and also for numerical estimation in generalised additive models (Hastie and Tibshirani 1990) with different smoothers not necessarily arising from a quadratic roughness functional.

2.3 The LMS approach to quantile regression

Although other methods of quantile regression are available, LMS, the semi-parametric approach pioneered by Tim Cole (Cole 1988), has a great deal of appeal, especially for the kinds of biometrical data arising in studies of childhood growth. His idea was to model the response Y_i at time t_i, perhaps the height or weight of a child aged t_i, as normally distributed after a Box–Cox transformation:

$$Z_i = \frac{(Y_i/\mu_i)^{\lambda_i} - 1}{\lambda_i \sigma_i} \sim N(0,1),$$

where the shape, median and spread parameters λ_i, μ_i and σ_i (hence, 'LMS') are smooth functions of t_i. In Cole (1988), a rather informal approach was taken to the smoothing required here, but in the RSS discussion of this paper I proposed fitting the same model using a more automatic/objective penalised likelihood approach, estimating functions $\mu(t)$, $\sigma(t)$ and $\lambda(t)$ by maximising

$$\ell(\mu,\sigma,\lambda) - \tfrac{1}{2}\alpha_\mu \int \mu''(t)^2 dt - \tfrac{1}{2}\alpha_\sigma \int \sigma''(t)^2 dt - \tfrac{1}{2}\alpha_\lambda \int \lambda''(t)^2 dt. \tag{2.4}$$

The log-likelihood ℓ takes the form

$$\sum_{i=1}^{n} \left(\lambda(t_i) \log \frac{Y_i}{\mu(t_i)} - \log \sigma(t_i) - \tfrac{1}{2} Z_i^2 \right),$$

where Z_i denotes Y_i after Box–Cox transformation as above. A more complete study of this idea was published subsequently as Cole and Green (1992), and these two papers have been widely cited and underpin growth reference curve methods used in practice world-wide.

The assumption of the normal distribution, after a quite-flexible smooth transformation, serves to stabilise inference about the distribution of Y give t across all quantiles, and the LMS method delivers estimates of all quantiles

$$\xi_\alpha(t) = \mu(t)\{1 + \lambda(t)\sigma(t)\Phi^{-1}(\alpha)\}^{1/\lambda(t)}$$

simultaneously, where Φ is the standard normal CDF. We thus have explicit 'plug-in' estimates for all quantile curves, and these come with a sanity-preserving guarantee not to touch or cross – features not shared by all competing methods. (Indeed none of the approaches covered in Kneib's paper enjoy both properties.) This essentially follows from the fact that the likelihood is derived from a model for the whole distribution of Y given x, and independently of the choice of quantile as the target of inference, in contrast to methods that have the formal structure of penalised likelihood, but in which the term corresponding to the likelihood itself is not in fact the probability density for the observed data. Again, we proposed back-fitting for numerical maximisation of the penalised log-likelihood (2.4). R code for fitting this model, and a related one based on gamma rather than normal distributions, was given by Marx (1999).

Rigby and Stasinopoulos (2005) took the final and important step of combining the ideas of transformation and generalised linear models, allowing on the way a fourth smooth curve to control kurtosis as well as skewness, location and spread, to create the rich and flexible class of GAMLSS models that are a major emphasis of the present paper; the Cole and Green (1992) LMS method is essentially the BCCG model in the GAMLSS class.

3 Looking forward

Since the days when penalised least squares and penalised likelihood were novel, the world of statistical model-fitting has been transformed by the blossoming of interest in Bayesian analysis. There are many reasons for this, and a reduction in scepticism about (or prejudice against) the Bayesian paradigm is only a small part of the story. The main drivers have been the extraordinary developments in Bayesian computation, through Markov chain Monte Carlo, and many other approaches, coupled with the huge increase in computer power routinely accessible to all statisticians, the theoretical developments in Bayesian analysis, in areas such as hierarchical models, nonparametric modelling and model determination, and above all the success stories, showing that in almost every conceivable domain of application, Bayesian methodology is delivering practical data analytic solutions even (and perhaps especially) in complex models.

How should we think about 'beyond mean regression' from a Bayesian perspective? Although Kneib takes 'Bayesian inference' as one of the three 'inferential procedures' that he concentrates on in Section 3, his discussion does not go much beyond a Bayesian-language re-interpretation of previously-mentioned methodologies. Few of the exciting and novel opportunities of taking a Bayesian approach to both modelling and inference are really unleashed. Let us start again.

Given covariate/response pairs (x_i, Y_i), we wish to make inference about the *conditional distribution* of Y given x – not only for x among the observed x_i, not only for $x \in \mathbb{R}^p$, not only for continuous Y, not only for the expectation of Y given x, not only *estimating* that conditional distribution, and not only for one x at a time but potentially for many x simultaneously, or even all x. That is, we postulate a family of distributions $\mathcal{F} = \{F_x, x \in \mathcal{X}\}$, where $Y|x \sim F_x$, assume that $Y_i \sim F_{x_i}$ independently (for the moment), and make inference about the family $\mathcal{F}$. We should, for example, be able to deliver the probability distribution of the set of x for which the median of F_x is positive – an unusual but not totally outlandish target, and one that is quite beyond the reach of a non-probabilistic analysis.

Such comprehensive inference about $\mathcal{F}$ is an ambitious goal indeed, and it would be naive to pretend that, without assumptions on structure, limitations on dimensions, etc., it could be universally practically attainable with real finite data sets. Nevertheless, Bayesian nonparametric regression at its most general seeks to address this goal, without making parametric assumptions on either the form of the distributions F_x or on the way these depend on x, in a fully probabilistic framework, that includes inference about the 'tuning constants' arising in a corresponding penalised likelihood set-up. This means that in principle it delivers the simultaneous posterior distribution $p(\mathcal{F}|(x_i, Y_i)_{i=1}^n)$; intrinsically, this is regression far 'beyond the mean'.

3.1 Bayesian nonparametrics

Bayesian nonparametric inference is not a brand-new phenomenon – modern practical methodologies have a 40-year line of descent running back to the important works of Ferguson (1973); Ferguson (1974). As with any other Bayesian model, we begin with a prior distribution on all unknowns – in the present case, that is a prior on $\mathcal{F}$. Much, but by no means all, modern Bayesian nonparametric analysis begins from a prior model on $\mathcal{F}$ that is based on the Dirichlet process (Ferguson 1973).

The Dirichlet process model for a single unknown distribution F is essentially the simplest natural generalisation to a general parameter space of the Beta prior on the parameter θ of a binary distribution $(1-\theta, \theta)$; this is the only prior on distributions that is conjugate to i.i.d. sampling. As well as this desirable property, which delivers obvious advantages for computation of posterior inference, the Dirichlet process enjoys strong theoretical properties, notably a richness of support that guarantees posterior consistency in several senses and settings. There are several equivalent definitions of the Dirichlet process, but the simplest to state is that a random distribution G on a space Ω follows a Dirichlet process $DP(\alpha, G_0)$ if for every partition $\{B_j\}_{j=1}^k$ of Ω, $(G(B_1), G(B_2), \ldots, G(B_k))$ has a Dirichlet distribution with parameters $(\alpha G_0(B_1), \alpha G_0(B_2), \ldots, \alpha G_0(B_k))$.

In most applied Bayesian methodology based on Dirichlet process models, the resulting random distribution G serves as the assumed model not directly for the observed responses Y_i, but for datum-specific parameters of which the Y_i are noisy versions. Such Dirichlet process mixture processes are (countably-) infinite mixture models that are both flexible and of great explanatory power. The weights on the components turn out to be generated a priori by a so-called 'stick-breaking' construction $\pi_h = V_h \prod_{l<h}(1-V_l)$, the V_l being i.i.d. Beta$(1, \alpha)$ variates.

For a comprehensive modern view on Bayesian nonparametrics, based on the Dirichlet process and otherwise, we refer the reader to Hjort *et al.* (2009).

3.2 Regression via density estimation

The simplest and earliest approach to using the Dirichlet process for nonparametric regression involves a clever trick, that takes us back to the original sense of the word 'regression' as meaning a conditional distribution in a bivariate/multivariate population. Müller *et al.* (1996) proposed to fit a Dirichlet process gaussian mixture model to $\{(x_i, Y_i)_{i=1}^n\}$ data, and then to deliver the conditional distribution of Y given x, exploiting the simple and explicit form this takes in a mutlivariate gaussian distribution. This analysis is available without programming through use of the `DPcdensity` function in the R package `DPpackage` (Jara *et al.* 2011).

Figure 1 shows some aspects of an analysis of the Munich rental data, including decile curves computed from the expected posterior conditional distribution function for Y (net rent) given x (area). As can be seen, this completely probabilistic analysis delivers conditional distributions that change smoothly and quite markedly with the covariate in location, spread and shape. This is just a simple example, and the methodology supports more elaborate modelling, including adjustment for other covariates.

In brief the model formulation here assumes that each data point (x_i, Y_i) is i.i.d. distributed according to a multivariate normal $N_{p+1}(\mu_i, \Sigma_i)$ say, $i = 1, 2, \ldots, n$, where the parameter pairs $\{(\mu_i, \Sigma_i)\}$ are drawn i.i.d. from a common distribution G given a Dirichlet process (α, G_0) prior. The concentration parameter α has a fixed Gamma prior, and the baseline distribution G_0 assumed to be Normal–inverse Wishart whose parameters in turn have weakly informative hyperpriors.

3.3 More recent methods

The analysis just described is a simple and convenient way of generating a very flexible analysis of regression data, but is subject to a criticism that can be important practically in the case that the covariate x is very irregularly distributed in the sample data. The Müller *et al.* (1996) approach essentially fits an infinite Normal–inverse Wishart model to the $\{(x_i, Y_i)_{i=1}^n\}$ data. Thus, irrespective of the distribution of Y given x in the data, if the covariate x is clustered, then the method tends to fit a large number of components, a kind of over-fitting through unnecessarily modelling the distribution of x; as a result, there will be diminished borrowing of strength between observations, adversely affecting quality of fit. To some extent, this issue arises even when x is more homogeneously distributed in the data, but is high-dimensional; again the model will focus more on modelling x than on Y given x (Wade *et al.* 2012).

There are a number of recently-proposed methods designed to avoid this difficulty; below I mention only a very small selection.

Dependent Dirichlet processes. MacEachern (1999) introduced a framework for very general nonparametric modelling of distributions P_x varying with covariates or spatial coordinates x. These are constructed by generalising the stick-breaking representation of a Dirichlet process mixture to allow the atoms (and sometimes also their weights) to depend on x.

A typical formulation used for regression would lead to a model for the density of Y given x of the form of a normal mixture

$$Y|x \sim \sum_{h=1}^{\infty} \pi_h N(y; \mu_h(x), \sigma_h^2)$$

where the weights are given by the usual stick-breaking formula, the μ_h independent draws from a suitable Gaussian process, and the variances σ_h^2 i.i.d. from an inverse gamma distribution.

Weighted mixture of DPs. Dunson *et al.* (2007) proposed a kernel-based approach to Bayesian density regression though use of a model

$$Y|x \sim \iint N(y; x_i^T \beta, \sigma^2) dP_x(\beta) d\pi(\sigma^2),$$

a mixture of multiple regression models with predictor-dependent weights. Their weighted mixture of Dirichlet processes prior models the x-dependent β distribution using Dirichlet bases placed at the sample predictor values:

$$P_x = \sum_{i=1}^{n} \left(\frac{\gamma_i K(x, x_i)}{\sum_{l=1}^{n} \gamma_l K(x, x_l)} \right) P_i^\star, \qquad P_i^\star \overset{\text{i.i.d.}}{\sim} DP(\alpha, P_0), \quad \forall x.$$

Kernel stick-breaking processes. Dunson and Park (2008) noted the unappealing sample-dependence in the specification of the mixing measures P_x, and proposed an alternative stick-breaking process, in which

$$P_x = \sum_{h=1}^{\infty} V_h K_{\psi_h}(x, \Gamma_h) \prod_{l<h} \{1 - V_l K_{\psi_l}(x, \Gamma_l)\} P_h^\star, \quad P_h^\star \overset{\text{i.i.d.}}{\sim} DP(\alpha, P_0)$$

where $V_h \overset{\text{i.i.d.}}{\sim} Beta(1, \lambda)$, K_ψ is a kernel with bandwidth ψ, $\{\psi_h\}$ a sequence of bandwidths sampled from G and $\{\Gamma_h\}$ a sequence of kernel locations sampled from H.

A predictive approach. Wade *et al.* (2012) deal with the difficulty mentioned at the beginning of the section by emphasising a predictive approach. The Dirichlet process mixture model of Müller *et al.* (1996) implicitly partitions the data by value of the x variable, and it is demonstrated that poor predictive performance can be blamed on the positive posterior probabilities assigned to 'bad' partitions. Even though these probabilities are very small, there are very many such partitions; Wade *et al.* (2012) specifically modify the marginal model for x to place zero prior probability on partitions that fail to respect a criterion of 'covariate proximity', and show convincingly improved performance.

3.4 Discussion

There is a useful review of many of these Bayesian nonparametric models for regression in the chapter by David Dunson in Hjort *et al.* (2009).

It is clear that this is a fast-moving field, and the topic is yet to mature. While all of these models support fully probabilistic inference about 'beyond-mean' regression, it can hardly be said that the approaches are fully comprehensible to the lay person. On the positive side, these models have large support, and consistency results are available. They are flexible, and the data-dependent character of the structure of the fitted models means that sparseness is built in, so that they automatically adjust when data follow a simple parametric law. Empirical performance has been studied up to a point. All of these methods can be implemented fairly efficiently but user-friendly software is not available, which limits both the take-up of the models and the extent to which they are understood and can be criticised. Above all, it is hard to argue that these are intuitive prior specifications in which the meaning and influence of hyperparameters is clear to the modeller and her client.

However, the direction of travel is to me clear, and I am confident that widely-acceptable methods of fully probabilistic regression will soon be available.

To conclude, I very much welcome Kneib's paper; he does us a great service with his emphasis on the importance of thinking of regression as modelling the *distribution* of responses given covariates, with the explanatory power and richness of interpretation that this can bring. I am sure that the next few years will see both more development of methodology and a better understanding of the advantages and disadvantages of each method; in the meantime, it is valuable to have this authoritative review of several current approaches.

Acknowledgement

I am very grateful to Alejandro Jara for making a copy of `DPpackage` version 1.1-7 available to me, in advance of its planned release in March 2013.

References

Breiman, L. and Friedman, J. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, **80**, (391), 580–98. (with discussion).

Buja, A., Hastie, T., and Tibshirani, R. (1989). Linear smoothers and additive models. *The Annals of Statistics*, **17**, (2), 453–510.

Cole, T. (1988). Fitting smoothed centile curves to reference data. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, **151**, 385–418.

Cole, T. J. and Green, P. J. (1992). Smoothing reference centile curves: the LMS method and penalized likelihood. *Statistics in Medicine*, **11**, 1305–19.

Dunson, D. and Park, J. (2008). Kernel stick-breaking processes. *Biometrika*, **95**, (2), 307–23.

Dunson, D., Pillai, N., and Park, J.-H. (2007). Bayesian density regression. *Journal of the Royal Statistical Society B*, **69**, 163–83.

Ferguson, T. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, **1**, 209–30.

Ferguson, T. (1974). Prior distributions on spaces of probability measures. *Annals of Statistics*, **2**, 615–29.

Green, P. J. (1987). Penalised likelihood for general semi-parametric regression models. *International Statistical Review*, **55**, 245–59.

Green, P. J., Jennison, C., and Seheult, A. H. (1985). Analysis of field experiments by least squares smoothing. *Journal of the Royal Statistical Society (B)*, **47**, 299–315.

Green, P. J. and Silverman, B. W. (1994). *Nonparametric regression and generalized linear models: a roughness penalty approach*. Chapman and Hall, London.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. Chapman and Hall.

Hjort, N., Holmes, C., Müller, P., and Walker, S. (2009). *Bayesian nonparametrics*. Cambridge University Press.

Jara, A., Hanson, T., Quintana, F., Müller, P., and Rosner, G. (2011). DPpackage: Bayesian semi- and nonparametric modeling in R. *Journal of Statistical Software*, **40**, (5), 1–30.

MacEachern, S. (1999). Dependent nonparametric process. *ASA Proceedings of the Section on Bayesian Statistical Science*.

Marx, B. D. (1999). Quantile regression (LMS) code. `http://www.stat.lsu.edu/faculty/marx/lms.txt/`.

Müller, P., Erkanli, A., and West, M. (1996). Bayesian curve fitting using multivariate normal mixtures. *Biometrika*, **83**, 67–79.

Rice, J. (1986). Convergence rates for partially splined models. *Statistics and Probability Letters*, **4**, (4), 203–208.

Rigby, R. A. and Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **54**, (3), 507–554.

Speckman, P. (1988). Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society. Series B (Methodological)*, **50**, 413–36.

Wade, S., Walker, S. G., and Petrone, S. (2012). A predictive study of Dirichlet process mixture models for curve fitting.

Wahba, G. (1990). *Spline models for observational data*. Society for Industrial and Applied Mathematics.

Wilkinson, G. N., Eckert, S. R., Hancock, T. W., and Mayo, O. (1983). Nearest neighbour (NN) analysis of field experiments. *Journal of the Royal Statistical Society. Series B (Methodological)*, **45**, (2), 151–211.

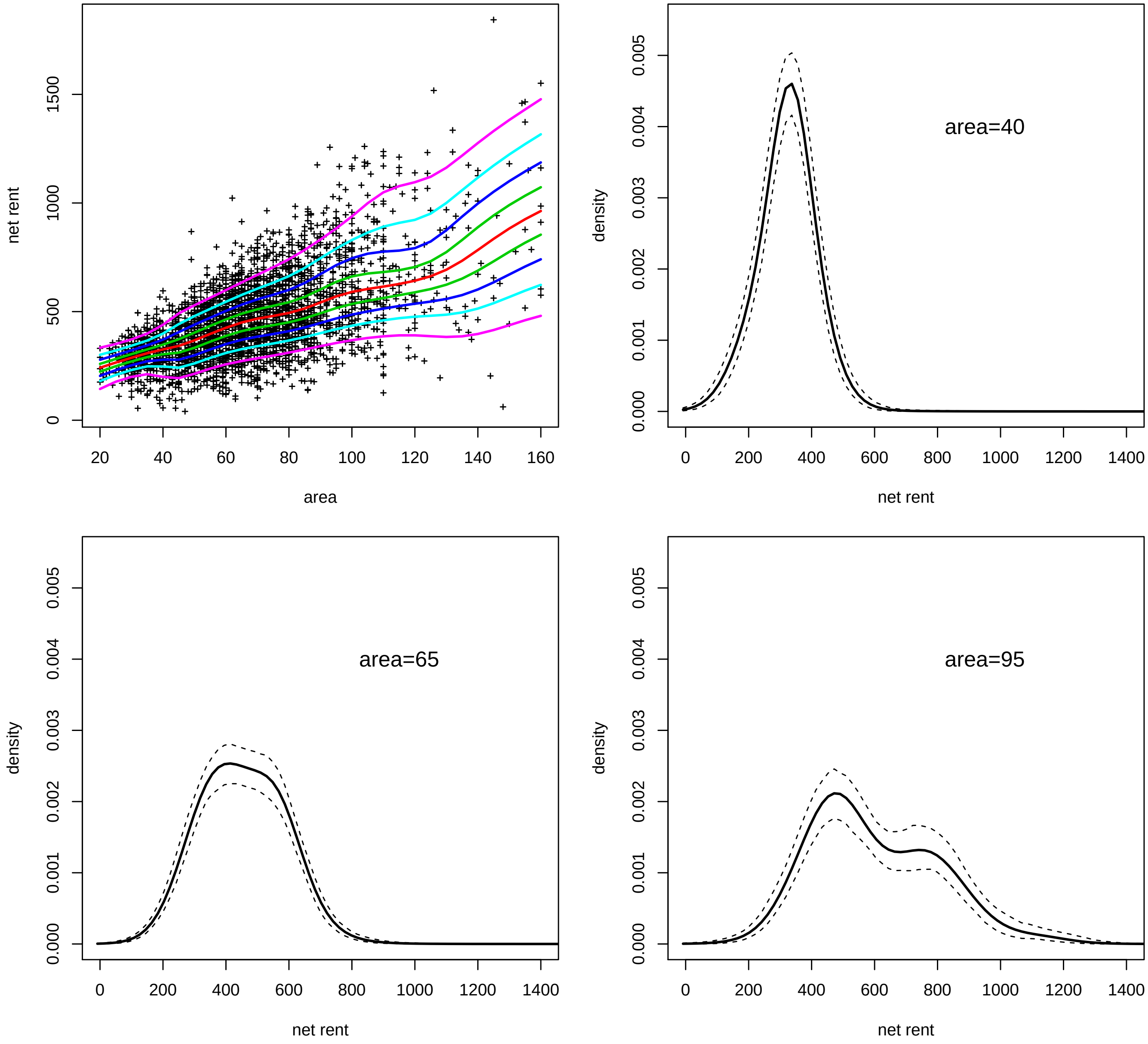

Figure 1: Munich rental guide data: a simple model fit using `DPcdensity`. Top left: raw data, net rent vs area, with 10%, 20%,. . . , 90% decile curves computed from the posterior mean of the cumulative conditional distribution of rent given area. Other panels: fitted conditional densities for rent, given area = 40, 65 and 95 respectively, with 90% credibility bands.