DENSITY ESTIMATION BY TOTAL VARIATION REGULARIZATION

ROGER KOENKER AND IVAN MIZERA

ABSTRACT. L_1 penalties have proven to be an attractive regularization device for nonparametric regression, image reconstruction, and model selection. For function estimation, L_1 penalties, interpreted as roughness of the candidate function measured by their total variation, are known to be capable of capturing sharp changes in the target function while still maintaining a general smoothing objective. We explore the use of penalties based on total variation of the estimated density, its square root, and its logarithm – and their derivatives – in the context of univariate and bivariate density estimation, and compare the results to some other density estimation methods including L_2 penalized likelihood methods.

Our objective is to develop a unified approach to total variation penalized density estimation offering methods that are: capable of identifying qualitative features like sharp peaks, extendible to higher dimensions, and computationally tractable. Modern interior point methods for solving convex optimization problems play a critical role in achieving the final objective, as do piecewise linear finite element methods that facilitate the use of sparse linear algebra.

1. INTRODUCTION

The appeal of pure maximum likelihood methods for nonparametric density estimation is immediately frustrated by the simple observation that maximizing log likelihoods,

$$\sum_{i=1}^{n} \log f(X_i) = \max_{f \in \mathcal{F}}!$$

over any moderately rich class of densities, $\mathcal{F}$, yields estimators that collapse to a sum of point masses. These notorious "Dirac catastrophes" can be avoided by penalizing the log likelihood

$$\sum_{i=1}^{n} \log f(X_i) - \lambda J(f) = \max_{f \in \mathcal{F}}! \tag{1}$$

by a functional J that imposes a cost on densities that are too rough. The penalty regularizes the original problem and produces a family of estimators indexed by the tuning parameter λ.

Version: February 14, 2006. This research was partially supported by NSF grant SES-02-40781, and by the Natural Sciences and Engineering Research Council of Canada.

Penalized maximum likelihood methods for density estimation were introduced by Good (1971), who suggested using Fisher information for the location parameter of the density as a penalty functional. Good offered a heuristic rationale of this choice as a measure of the sensitivity of the density to location shifts. The choice has the added practical advantage that it permits the optimization to be formulated as a convex problem with the (squared) L_2 penalty,

$$J(f) = \int (\sqrt{f}\,')^2 dx. \tag{2}$$

In subsequent work Good and Gaskins (1971) found this penalty somewhat unsatisfactory, producing estimates that sometimes "looked too straight." They suggested a modified penalty that incorporated a component penalizing the *second* derivative of $\sqrt{f}$ as well as the first. This component has a more direct interpretation as a measure of curvature and therefore as a measure of roughness of the fitted density.

Eschewing a "full-dress Bayesian approach," Good and Gaskins refer to their methods as a "Bayesian approach in mufti." Ideally, penalties could be interpreted as an expression of prior belief about the plausibility of various elements of $\mathcal{F}$. In practice, the justification of particular penalties inevitably has a more heuristic, ad-hoc flavor: data-analytic rationality constrained by computational feasibility. While penalties may be applied to the density itself rather than to its square root, a possibility briefly mentioned in Silverman (1986), a more promising approach considered by Leonard (1978) and Silverman (1982) replaces $\sqrt{f}$ by $\log f$ in the penalty term. When the second derivative of $\log f$ is penalized, this approach privileges exponential densities; whereas penalization of the third derivative of $\log f$ targets the normal distributions.

The early proposals of Good and Good and Gaskins have received detailed theoretical consideration by Thompson and Tapia (1990) and by Eggermont and LaRiccia (2001), who establish consistency and rates of convergence. A heuristic argument of Klonias (1991) involving influence functions suggests that penalized likelihood estimators perform automatically something similar in effect to the "data sharpening" of Hall and Minnotte (2002) – they take mass from the "valleys" and distribute it to the "peaks." Silverman (1984) provides an nice link between penalty estimators based on the rth derivative of $\log f$ and adaptive kernel estimators, and he suggests that the penalty approach achieves a degree of automatic adaptation of bandwidth without reliance on a preliminary estimator. Taken together this work constitutes, we believe, a convincing *prima facie* case for the regularization approach to density estimation.

From the computational point of view, all these proposals, starting from those of Good, can be formulated as convex optimization problems and therefore are in principle efficiently computable. However, the practice has not been that straightforward, and widely accessible implementations may still not be always available. In particular, the earlier authors thinking in terms of techniques for minimization of quadratic functionals might still view the constraints implied by the fact that the optimization

must be performed over f that are densities as a computational pain. Penalization of $\sqrt{f}$ or $\log f$ is often motivated as a practical device circumventing the nonnegativity constraint on f; penalizing the logarithm of the density as noted by Silverman (1982), offers a convenient opportunity to eliminate the constraint requiring the integral of f to be 1. In contrast to these advantages, penalizing the density f itself requires a somewhat more complicated strategy to ensure the positivity and integrability of the estimator. In any case, the form of the likelihood keeps the problem nonlinear; hence iterative methods are ultimately required. Computation of estimators employing the L_2 penalty on $(\log f)''$ has been studied by O'Sullivan (1988). An implementation in R is available from the package `gss` of Gu (2005). Silverman's (1982) proposal to penalize the third derivative of $\log f$, thereby shrinking the estimate toward the Gaussian density, has been implemented by Ramsay and Silverman (2002).

The development of modern interior-point methods for convex programming not only changes this outlook – convex programming works with constraints routinely – but also makes various other penalization proposals viable. In what follows, we would like to introduce several new nonparametric density estimation proposals involving penalties formulated in terms of total variation. Weighted sums of squared L_2 norms are replaced by weighted L_1 norms as an alternative regularization device. Squaring penalty contributions inherently exaggerates the contribution to the penalty of jumps and sharp bends in the density; indeed, density jumps and piecewise linear bends are impossible in the L_2 framework since the penalty evaluates them as "infinitely rough." Total variation penalties are happy to tolerate such jumps and bends, and they are therefore better suited to identifying discrete jumps in densities or in their derivatives. This is precisely the property that has made them attractive in imaging applications.

From a computational perspective, total-variation based penalties fit comfortably into modern convex optimization setting. Exploiting the inherent sparsity of the linear algebra required yields very efficient interior point algorithms. We will focus our attention on penalizing derivatives of $\log f$, but other convex transformations can be easily accommodated. Our preliminary experimentation with penalization of $\sqrt{f}$ and f itself did not seem to offer tangible benefits.

Total-variation penalties also offer natural multivariate generalizations. Indeed, we regard univariate density estimation as only a way station on a road leading to improved multivariate density estimators. To this end, the fact that penalty methods can easily accommodate qualitative constraints on estimated functions and their boundary values is an important virtue. One of our original motivations for investigating total variation penalties for density estimation was the ease with which qualitative constraints – monotonicity or log-concavity, for instance – could be imposed. In this context it is worth mentioning the recent work of Rufibach and Dümbgen (2004) who show that imposing log-concavity *without any penalization* is enough to regularize the univariate maximum likelihood estimator, and achieve attractive asymptotic behavior.

Total variation penalties for nonparametric regression with scattered data have been explored by Koenker, Ng, and Portnoy (1994), Mammen and van de Geer (1997), Davies and Kovac (2001, 2004) and Koenker and Mizera (2002, 2004). Total variation has also played an important role in image processing since the seminal papers of Mumford and Shah (1989), and Rudin, Osher, and Fatemi (1992).

We begin by considering the problem of estimating univariate densities, and then extend the approach to bivariate settings.

2. Univariate Density Estimation

Given a random sample, $X_1, \ldots, X_n$ from a density f_0, we will consider estimators that solve,

$$\max_f \{\sum_{i=1}^n \log f(X_i) - \lambda J(f) \mid \int_\Omega f = 1\}, \tag{3}$$

where J denotes a functional intended to penalize for the roughness of candidate estimates, $\mathcal{F}$, and λ is a tuning parameter controlling the smoothness of the estimate. Here the domain Ω may depend on *a priori* considerations as well as the observed data.

We propose to consider roughness penalties based on total variation of the transformed density and its derivatives. Recall that the total variation of a function $f : \Omega \to \mathcal{R}$ is defined as

$$\bigvee_\Omega (f) = \sup \sum_{i=1}^m |f(u_i) - f(u_{i-1})|,$$

where the supremum is taken over all partitions, $u_1 < \ldots < u_m$ of Ω. When f is absolutely continuous, we can write, see e.g. Natanson (1974, p.259),

$$\bigvee_\Omega (f) = \int_\Omega |f'(x)| dx.$$

We will focus on penalizing the total variation of the first derivative of the log density,

$$J(f) = \bigvee_\Omega ((\log f)') = \int_\Omega |(\log f)''|,$$

so letting $g = \log f$ we can rewrite (3) as,

$$\max_g \{\sum_{i=1}^n g(X_i) - \lambda \bigvee_\Omega (g') \mid \int_\Omega e^g = 1\}. \tag{4}$$

But this is only one of many possibilities: one may consider

$$J(f) = \bigvee_\Omega (g^{(k)}),$$

where $g^{(0)} = g$, $g^{(1)} = g'$, etc., and g may be $\log f$, or $\sqrt{f}$, or f itself, or more generally $g^\kappa = f$, for $\kappa \in [1, \infty]$, with the convention that $g^\infty \equiv e^g$. Even more generally, linear combinations of such penalties with positive weights may be considered. From this family we adopt $\kappa = \infty$ and $k = 1$; see Sardy and Tseng (2005) for $\kappa = 1$ and $k = 0$. In multivariate settings $g^{(k)}$ is replaced by $\nabla^k g$, as described in the next section.

As noted by Gu (2002), even for L_2 formulations the presence of the integrability constraint prevents the usual reproducing kernel strategy from finding exact solutions; some iterative algorithm is needed. We will adopt a finite element strategy that enables us to exploit the sparse structure of the linear algebra used by modern interior point algorithms for convex programming.

Restricting attention to f's for which $\log(f)$ is piecewise linear on a specified partition of Ω, we can write $J(f)$ as an ℓ_1 norm of the second weighted differences of f evaluated at the mesh points of the partition. More explicitly, let Ω be the closed interval $[x_0, x_m]$ and consider the partition $x_0 < x_1 < \cdots < x_m$ with spacings $h_i = x_i - x_{i-1}, \quad i = 1, \cdots m$. If $\log(f(x))$ is piecewise linear, so that

$$\log(f(x)) = \alpha_i + \beta_i x \qquad x \in [x_i, x_{i+1}),$$

then

$$J(f) = \bigvee_\Omega((\log f)') = \sum_{i=1}^m |\beta_i - \beta_{i-1}| = \sum_{i=1}^m |(\alpha_{i+1} - \alpha_i)/h_{i+1} - (\alpha_i - \alpha_{i-1})/h_i|,$$

where we have imposed continuity of f in the last step. We can thus parameterize functions $f \in \mathcal{F}$ by the function values $\alpha_i = \log(f(x_i))$, and this enables us to write our problem (3) as a linear program,

$$\max\{\sum_{i=1}^n \alpha_i - \lambda \sum_{j=1}^m (u_j + v_j) | D\alpha - u + v = 0,\ (\alpha, u, v) \in \mathbb{R}^n \times \mathbb{R}_+^{2m}\} \tag{5}$$

where D denotes a tridiagonal matrix containing the h_i factors for the penalty contribution, and u and v represent the positive and negative parts of the vector $D\alpha$, respectively.

An advantage of parameterization of the problem in terms of $\log f$ is that it obviates any worries about the non-negativity of $\hat{f}$. But we have still neglected one crucial constraint. We need to ensure that our density estimates integrate to one. In the piecewise linear model for $\log f$ this involves a rather awkward nonlinear constraint on the α's,

$$\sum_{j=1}^m h_i \frac{e^{\alpha_i} - e^{\alpha_{i-1}}}{\alpha_i - \alpha_{i-1}} = 1.$$

This form of the constraint cannot be incorporated directly in its exact form into our optimization framework, nevertheless its approximation by a Riemann sum on a sufficiently fine grid provides a numerically satisfactory solution.

2.1. **Data Augmentation.** In the usual Bayesian formalism, the contribution of the prior can often be represented as simple data augmentation. That is, the prior can be interpreted as what we would believe about the model parameters if we had observed some "phantom data" whose likelihood we could evaluate. This viewpoint may strain credulity somewhat, but under it the penalty, $J(f)$, expresses the belief that we have "seen" m observations on the second differences of $\log f$ evaluated at the x_i's, *all zero*, and independent with standard Laplacian density, $\frac{1}{2}e^{-|x|}$. The presence of λ introduces a free scale parameter that represents the strength of this belief. Data dependent strategies for the choice of λ obviously violate Bayesian orthodoxy, but maybe condoned by the more pragmatic minded.

Pushing the notion of Bayesian virtual reality somewhat further, we may imagine observing data at new x_i values. Given that our estimated density is parameterized by its function values at the "observed" x_i values, these new values introduce new parameters to be estimated; these "phantom observations" contribute nothing to the likelihood, but they do contribute to the penalty term $J(f)$. But by permitting $\log f$ to bend at the new x_i points in regions where there is otherwise no real data, flexibility of the fitted density is increased. In regions where the function $\log f$ is convex, or concave, one large change in the derivative can thus be broken up into several smaller changes, without affecting the total variation of its derivative. Recall that the total variation of a monotone function on an interval is just the difference in the values taken at the endpoints of the interval.

Rather than trying to carefully select a few x_i values as knots for a spline representation of the fitted density, as described in Stone, Hansen, Kooperberg, and Truong (1997), all of the observed x_i are retained as knots and some virtual ones are thrown in as well. Shrinkage, controlled by the tuning parameter, λ, is then relied upon to achieve the desired degree of smoothing. The use of virtual observations is particularly advantageous in the tails of the density, and in other regions where the observed data are sparse. We will illustrate the use of this technique in both univariate and bivariate density estimation in the various examples of subsequent sections.

Example. Several years ago one of us, as a class exercise, asked students to estimate the density illustrated in Figure 1(a), based on a random sample of 200 observations. The density is a mixture of three, three-parameter lognormals:

$$f_1(x) = \sum_{i=1}^{3} w_i \phi(\log((x - \gamma_i - \mu_i)/\sigma_i))/(\sigma_i(x - \gamma_i)), \tag{6}$$

where ϕ denotes the standard normal density, $\mu = (0.5, 1.1, 2.6)$, $\gamma = (.0.4, 1.2, .2.4)$, $\sigma = (0.2, 0.3, .0.2)$, and $w = (0.33, 0.33, 0.33)$. In the figure we have superimposed the density on a histogram of the original data using an intentionally narrow choice of binwidth.

The most striking conclusion of the exercise was how poorly conventional density estimators performed. With one exception, none of the student entries in the

competition were able to distinguish the two tallest peaks, and their performance on the lower peak wasn't much better. All of the kernel estimates looked very similar to smoother of the two kernel estimates displayed in Figure 1(b). This is a fixed-bandwidth Gaussian kernel estimate with bandwidth chosen by Scott's (1992) biased cross-validation criterion as implemented in R and described by Venables and Ripley (2002). The other kernel estimate employs Scott's alternative unbiased cross-validation bandwidth, and clearly performs somewhat better. Gallant and Nychka's (1987) Hermite series estimator also oversmooths when the order of the estimator is chosen with their BIC criterion, but performs better when AIC order selection is used, as illustrated in Figure 1(c). In Figure 1(d) we illustrate two variants of the most successful of the student entries based on the logspline method of Kooperberg and Stone (1991): one constrained to have positive support, the other unconstrained. Figure 1(e) illustrates two versions of the logspline estimator implemented by Gu (2002). Finally, Figure 1(f) illustrates two versions of a total variation penalty estimator; both versions employ a total variation penalty on the derivative of $\log f$, and use in addition to the 200 sample observations, 300 "virtual observations" equally spaced between 0 and 25. These estimators were computed with the aid of the MOSEK package of E. D. Andersen, an implementation for MATLAB of the methods described in Andersen and Ye (1998). The penalty method estimators all perform well in this exercise, but the kernel and Hermite series estimators have difficulty coping with the combination of sharp peaks and smoother foothills.

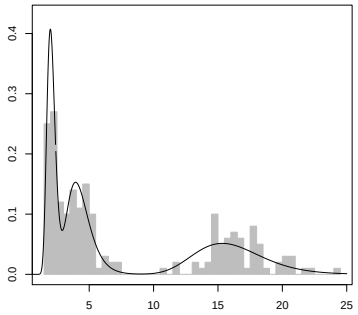
3. Bivariate Density Estimation

In nonparametric regression piecewise linear fitting is often preferable to piecewise constant fitting. Thus, penalizing total variation of the gradient, ∇g, instead of total variation of g itself, is desirable. For smooth functions we can extend the previous definition by writing,

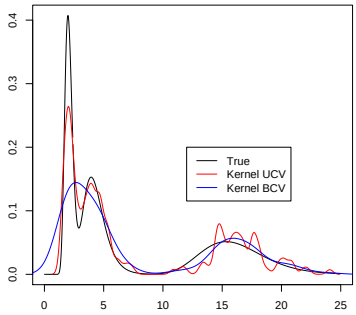
$$\bigvee_{\Omega} \nabla g = \int_{\Omega} \|\nabla^2 g\|, \tag{7}$$

where $\|\cdot\|$ can be taken to be the Hilbert-Schmidt norm, although other choices are possible as discussed in Koenker and Mizera (2004). This penalty is closely associated with the thin plate penalty that replaces $\|\nabla^2 g\|$ with $\|\nabla^2 g\|^2$. The latter penalty has received considerable attention, see e.g. Wahba (1990) and the references cited therein. We would stress, however, that as in the univariate setting there are important advantages in taking the square root.

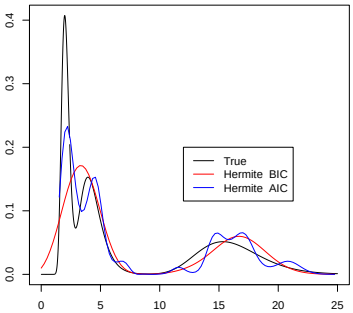
For scattered data more typical of nonparametric regression applications, Koenker and Mizera (2004) have proposed an alternative discretization of the total variation penalty based on continuous, piecewise-linear functions defined on triangulations of a convex, polyhedral domain. Following Hansen, Kooperberg, and Sardy (1998), such functions are called triograms. The penalty (7) can be simplified for triograms by



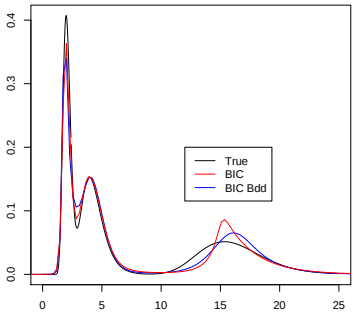
(a) Histogram



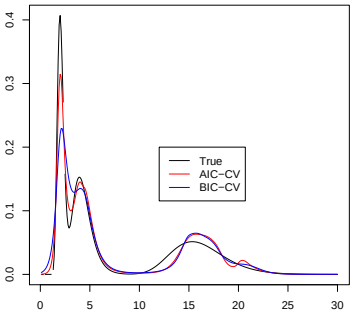
(b) Kernel



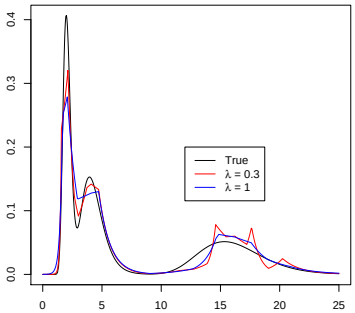
(c) Hermite



(d) Logspline



(e) Gulog



(f) TVlog

FIGURE 1. Comparison of Estimates of the 3-Sisters Density.

summing the contributions over the edges of the triangulation,

$$\bigvee_{\Omega} \nabla g = \sum_k \|\nabla g^+_{e_k} - \nabla g^-_{e_k}\| \, \|e_k\|. \tag{8}$$

Each edge is associated with two adjacent triangles; the contribution of the edge is simply the product of the Euclidean norm of the difference between the gradients on the two triangles multiplied by the length of the edge. The interiors of the triangles, since they are linear, contribute nothing to the total variation, nor do the vertices of the triangulation. See Koenker and Mizera (2004) for further details.

Choice of the triangulation is potentially an important issue especially when the number of vertices is small, but numerical stability favors the classical Delaunay triangulation in most applications. Hansen and Kooperberg (2002) consider sequential (greedy) model selection strategies for choosing a parsimonious triangulations for nonparametric regression without relying on a penalty term. In contrast, Koenker and Mizera (2004) employ the total variation penalty (8) to control the roughness of the fit based on a much more profligate triangulation. As in the univariate setting it is often advantageous to add virtual vertices that can improve the flexibility of the fitted function.

Extending the penalized triogram approach to bivariate density estimation requires us, as in the univariate case, to make a decision about what is to be penalized? We will focus exclusively on total variation penalization of the log density with the understanding that similar methods could be used for the density itself or another (convex) transform.

Given independent observations $\{x_i = (x_{1i}, x_{2i}) : i = 1, \cdots, n\}$ from a bivariate density $f(x)$, let $g = \log f$, and consider the class of penalized maximum likelihood estimators solving

$$\max_{g \in \mathcal{G}} \sum_{i=1}^{n} g(x_i) - \lambda J(g),$$

where J is the triogram penalty, given by (8). The set $\mathcal{A}$ consists of triogram densities: continuous functions from a polyhedral convex domain Ω to $\mathbb{R}_+$, piecewise linear on a triangulation of Ω and satisfying the condition,

$$\int_{\Omega} e^g = 1.$$

It follows that $\log f$ can be parameterized by its function values at the vertices of the triangulation. As in the univariate case, adding virtual vertices is advantageous especially so in the region outside the convex hull of the observed data where they provide a device to cope with tail behavior.

Example. To illustrate the performance of our bivariate density estimator, we consider the density

$$\begin{aligned} f_2(x_1, x_2) &= f(x_2|x_1)f(x_1) \\ &= 2\phi(2(x_2 - \sqrt{x_1})) \cdot f_1(x_1), \end{aligned}$$

where f_1 is the univariate test density given above. Two views of this density can be seen in the upper panels of Figure 2. There is one very sharp peak and two narrow "fins". In the two lower panels we depict views of a fitted density based on 1000 observations. The tuning parameter λ is taken to be 2, and the fit employs virtual observations on a integer grid over the rectangle $\{[0, 30] \times [0, 6]\}$.

4. Duality and Regularized Maximum Entropy

An important feature of convex optimization problems is that they may be reformulated as dual problems, thereby often offering a complementary view of the problem from the other side of the looking glass. In addition to providing deeper insight into the interpretation of the problem as originally posed, dual formulations sometimes yield substantial practical benefits in the form of gains in computational efficiency. In our experience, the dual formulation of our computations exhibits substantially better performance than the original penalized likelihood formulation. Execution times are about 20 percent faster and convergence is more stable. We will show in this section that total variation penalized maximum likelihood density estimation has a dual formulation as regularized form of maximum entropy estimation.

As we have seen already, piecewise linear log density estimators can be represented by a finite dimensional vector of function values

$$\alpha_i = g(x_i) \quad i = 1, \cdots, m,$$

evaluated at knot locations, $x_i \in \Omega$. These points of evaluation can be sample observations or "virtual" observations, or a mixture of the two. They may be univariate, bivariate, or in principle, higher dimensional. We approximate our integral by the Riemann sum,

$$\int_\Omega e^g \approx \sum_{i=1}^m c_i e^{\alpha_i},$$

a step that can be justified rigorously by introducing points of evaluation on a sufficiently fine grid, but is also motivated by computational considerations. Provisionally, we will set the tuning parameter $\lambda = 1$, so our primal problem is,

$$\max\{\delta^\top \alpha - \|D\alpha\|_1 \quad | \ \sum_i c_i e^{\alpha_i} = 1\}. \tag{P}$$

In the simplest case the vector $\delta \in \mathbb{R}$ is composed of zeros and ones indicating which elements of α correspond to sample points and thus contribute to the likelihood term. In the case that the x_i are *all* virtual, chosen to lie on a regular grid, for example,

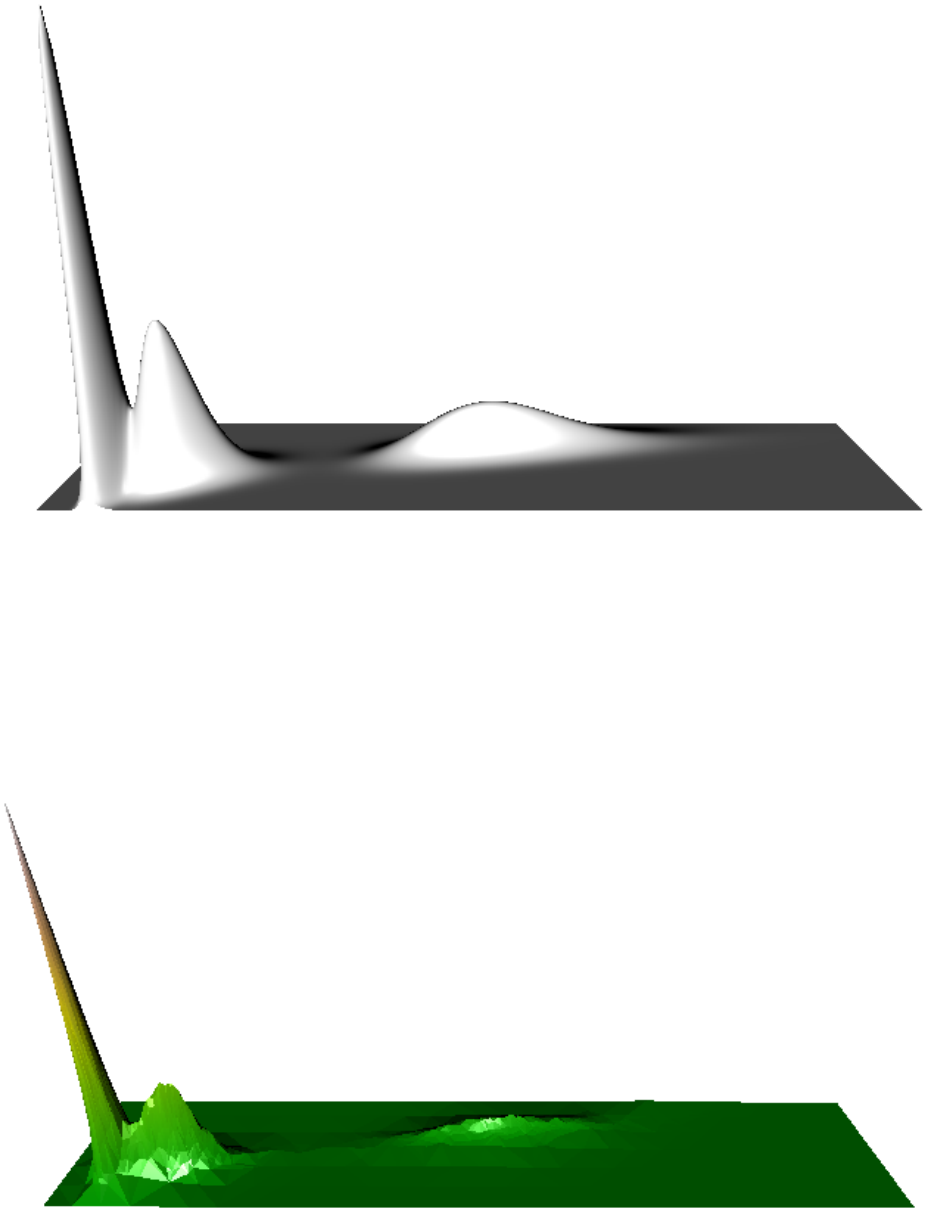


Figure 2. Bivariate 3-Sisters Density and an Estimate.

we can write, $\delta = B1_n$, where B denotes an m by n matrix representing the n sample observations expressed in terms of the virtual points, e.g. using barycentric coordinates.

The integrability constraint can be conveniently incorporated into the objective function using the following discretized version of a result of Silverman (1982).

Lemma 1. *$\hat{\alpha}$ solves problem (P) if and only if $\hat{\alpha}$ maximizes,*

$$R(\alpha) = \delta^\top \alpha - \|D\alpha\|_1 - n \sum_i c_i e^{\alpha_i}.$$

Proof. Note that any differential operator, D, annihilates constant functions, or the vector of ones. Thus, evaluating R at $\alpha^* = \alpha - \log \sum c_i e^{\alpha_i}$, so $\sum c_i e^{\alpha_i^*} = 1$, we have

$$R(\alpha^*) = R(\alpha) + n \sum_i c_i e^{\alpha_i} - n \log \sum_i c_i e^{\alpha_i} - 1,$$

but $t - \log t \geq 1$, for all $t > 0$ with equality only at $t = 1$. Thus, $R(\alpha^*) \geq R(\alpha)$, and it follows that $\hat{\alpha}$ maximizes R if and only if $\hat{\alpha}$ maximizes R subject to $\sum_i c_i e^{\alpha_i} = 1$. This constrained problem is equivalent to (P). ∎

Introducing the artificial barrier vector β, the penalty contribution can be reformulated slightly, and we can write (P) as,

$$\max_{\alpha,\beta} \{\delta^\top \alpha - 1^\top \beta - \sum_i c_i e^{\alpha_i} \mid D\alpha \leq \beta, \quad -D\alpha \leq \beta\}.$$

We seek to minimize the Lagrangian,

$$\begin{aligned} L(\alpha, \beta, \nu_1, \nu_2) &= \delta^\top \alpha - 1^\top \beta - n \sum c_i e^{\alpha_i} + \nu_1^\top (D\alpha - \beta) + \nu_2^\top (-D\alpha - \beta) \\ &= (\delta + D^\top (\nu_1 - \nu_2))^\top \alpha - (1 - \nu_1 - \nu_2) 1^\top \beta - n \sum c_i e^{\alpha_i}, \end{aligned}$$

subject to the feasibility constraints,

$$\gamma \equiv \delta + D^\top (\nu_1 - \nu_2) \geq 0, \ \nu_1 + \nu_2 = 1, \ \nu_1 \geq 0, \text{ and } \nu_2 \geq 0.$$

Now, differentiating the Lagrangian expression with respect to the α_i's yields

$$\frac{\partial L}{\partial \alpha_i} = \delta_i - d_i^\top (\nu_1 - \nu_2) - c_i e^{\alpha_i} = 0, \ i = 1, \cdots, m.$$

Convexity assures that these conditions are satisfied at the unique optimum:

$$f_i \equiv (\delta_i - d_i^\top (\nu_1 - \nu_2))/c_i = e^{\alpha_i} \quad i = 1, \cdots, m,$$

so we can rewrite our Lagrangian problem with $C = \text{diag}(c)$ as

$$\min\{\sum c_i f_i \log f_i \mid f = C^{-1}(\delta + D^\top y) \geq 0. \ \|y\|_\infty \leq 1\}.$$

Reintroducing the tuning parameter λ we obtain the final form of the dual problem.

Theorem 1. *Problem (P) has equivalent dual formulation*

$$\max\{-\sum c_i f_i \log f_i \mid f = C^{-1}(\delta + D^\top y) \geq 0,\ \|y\|_\infty \leq \lambda\}. \tag{D}$$

Remarks:

(1) We can interpret the dual as a maximum entropy problem regularized by the ℓ_∞ constraint on y with added requirement that an affine transformation of the vector of dual variables, y, lies in the positive orthant.

(2) The ℓ_∞ constraint may be viewed as a generalized form of the tube constraint associated with the taut string methods of Davies and Kovac (2004). In the simplest setting, when total variation of the log density itself, rather than its derivative, is employed as a penalty for univariate density estimation, D is a finite difference operator and the dual vector, y, can be interpreted as a shifted estimate of the distribution function constrained to lie in a band around the empirical distribution function. In more general settings the geometric interpretation of the constraints on the dual vector, y, in terms of the sample data is somewhat less clear.

(3) The weights c_i appearing in the objective function indicate that the sum may be interpreted as a Riemann approximation to the entropy integral. Expressing the problem equivalently as the maximization of

$$\sum_i c_i f_i \log \frac{c_i}{c_i f_i} + \log n$$

we arrive at an interpretation in terms of the Kullback-Leibler divergence, $\mathcal{K}(\phi, \nu)$, of the probability distribution $\phi = (c_i f_i)$, corresponding to the estimated density f, from the probability distribution $\nu = n(c_i)$, corresponding to the density uniform over Ω. Thus, our proposal can be interpreted in terms of regularized minimum distance estimation,

$$\min\{\mathcal{K}(\phi, \nu) | \phi = (\delta + D^T y) \geq 0, \|y\|_\infty < \lambda\},$$

a formulation not entirely surprising in the light of our knowledge about maximum likelihood estimation. The choice of the uniform "carrier" density could be modified to obtain exponentially tilted families as described in Efron and Tibshirani (1996).

(4) Density estimation methods based on maximum entropy go back at least to Jaynes (1957). However, this literature has generally emphasized imposing exact moment conditions, or to use the machine learning terminology, "features," on the estimated density. In contrast, our dual problem may be viewed as a regularized maximum entropy approach that specifies "soft" feature constraints imposed as inequalities. Dudík, Phillips, and Schapire (2004) consider a related maximum entropy density estimation problem with soft feature constraints. Donoho, Johnstone, Hoch, and Stern (1992) consider related penalty

methods based on entropy for a class of regression type imaging and spectroscopy problems, and show that they have superior performance to linear methods based on Gaussian likelihoods and priors.

5. MONTE-CARLO

In this section we report the results of a small Monte-Carlo experiment designed to compare the performance of the TV penalized estimator with three leading competitors:

TS: The taut string estimator of Davies and Kovac (2005) using the default tuning parameters embedded in the function `pmden` of their R package `ftnonpar`.

Kucv: The fixed bandwidth kernel density estimator implemented by the function `density` in the R `stats` package, employing Scott's (1992) "unbiased cross validation" bandwidth selection.

Kbcv: The fixed bandwidth density estimator as above, but using Scott's biased cross-validation bandwidth.

For purposes of the Monte-Carlo, automatic selection of λ for the TV estimator was made according to the following recipe. Estimates were computed at the fixed λ's, $\{.1, .2, \ldots, .9, 1.0\}$, using virtual observations on a grid, $\mathcal{G}$, of 400 points equally spaced on $[-4, 4]$. For each of these estimates the Kolmogorov distance between the empirical distribution function of the sample, $\hat{F}_n$, and the smoothed empirical, $\tilde{F}_{n,\lambda}$, corresponding to the density estimate

$$\kappa(\lambda) \equiv K(\hat{F}_n, \tilde{F}_{n,\lambda}) = \max_{x_i \in \mathcal{G}} |\hat{F}_n(x_i) - \tilde{F}_{n,\lambda}(x_i)|$$

was computed. Based on preliminary investigation, $\log \kappa(\lambda)$ was found to be approximately linear in $\log \lambda$, so we interpolated this log-linear relationship to find the λ that made $\kappa(\lambda)$ approximately equal to the cutoff $c_\kappa = .3/\sqrt{n}$. The value .3 was chosen utterly without any redeeming theoretical justification. In rare cases for which this interpolation fails, i.e., $\hat{\lambda} \notin [.1, 1]$, we use $\hat{\lambda} = \max\{\min\{\hat{\lambda}, 1\}, .1\}$.

As candidate densities, we use the familiar Marron and Wand (1992) normal mixtures illustrated in Figure 1. Random samples from these densities were generated in with the aid of the R `nor1mix` package of Mächler (2005). All computations for the taut string and kernel estimators are conducted in R; computations for the TV estimator are made in matlab with the aid of the `PDCO` function of Saunders (2004) as described above using the sample data generated from R.

Three measures of performance are considered for each of the 16 test densities. Table 1.1 reports the proportion replications for which each method obtained the correct identification of the number of modes of the true density. Table 1.2 reports median MIAE (mean integrated absolute error), and Table 1.3 reports median MISE (mean integrated squared error).

Clearly, the taut-string estimator performs very well in identifying unimodal and well separated bimodal densities, but it has more difficulties with the multimodal

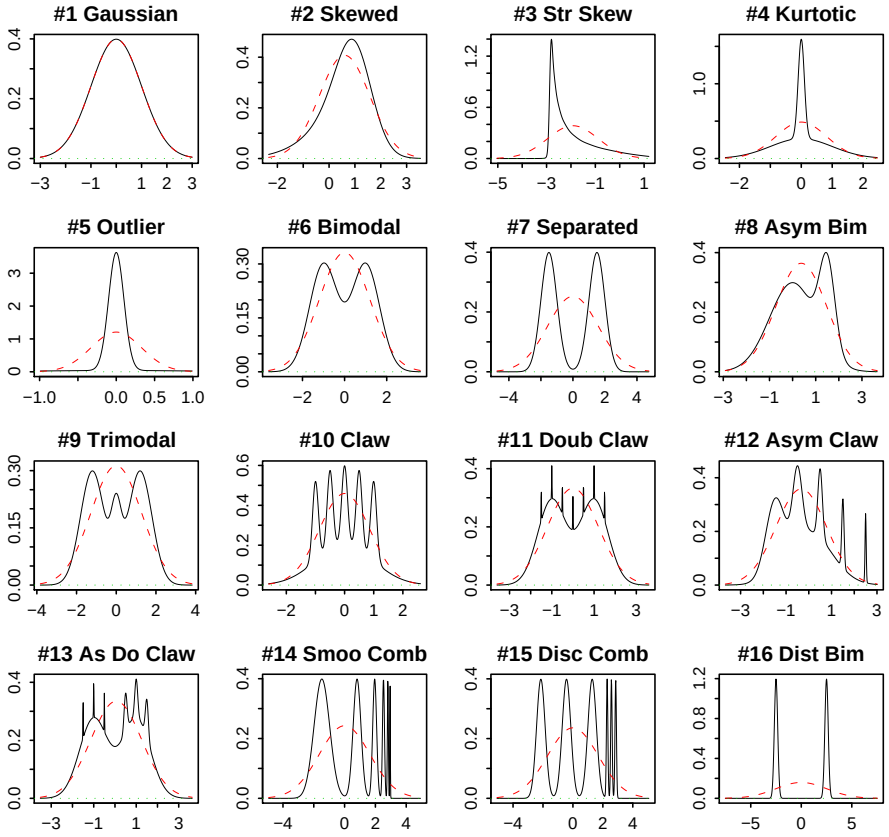


FIGURE 3. The Marron and Wand candidate densities.

cases. Unbiased cross-validation is generally inferior to biased cross-validation from a mode identification viewpoint, producing too rough an estimate and therefore too many modes.

Unbiased CV has quite good MIAE performance. Not surprisingly, it does best at the normal model, but it is somewhat worse than our TV estimator for distributions 3, 4, 5, and 16. In the other cases the performance is quite comparable. The biased CV kernel estimator is consistently inferior in MIAE except at the normal model. It fails spectacularly for the sharply bimodel density number 16. The TV estimator is not too bad from the MIAE perspective, consistently outperforming the taut-string estimator by a substantial margin, and very competitive with the kernel estimators except in the strictly Gaussian setting. Results for MISE are generally similar to those for MIAE.

Distribution	TV	TS	K-ucv	K-bcv
MW 1	0.303	1.000	0.690	0.863
MW 2	0.304	1.000	0.354	0.456
MW 3	0.169	1.000	0.000	0.059
MW 4	0.152	1.000	0.000	0.176
MW 5	0.345	1.000	0.000	0.000
MW 6	0.634	0.329	0.718	0.937
MW 7	0.716	1.000	0.678	0.880
MW 8	0.522	0.067	0.279	0.592
MW 9	0.472	0.013	0.434	0.292
MW 10	0.680	0.528	0.000	0.001
MW 11	0.008	0.000	0.006	0.000
MW 12	0.438	0.014	0.017	0.000
MW 13	0.016	0.001	0.003	0.000
MW 14	0.056	0.021	0.000	0.014
MW 15	0.078	0.078	0.000	0.038
MW 16	0.914	1.000	0.000	1.000

TABLE 1. Proportion of correct estimates of the number of modes: Sample size, $n = 500$ and number of replications $R = 1000$.

6. Prospects and Conclusions

Total variation penalty methods appear to have some distinct advantages when estimating densities with sharply defined features. They also have attractive computational features arising from the convexity of the penalized likelihood formulation.

There are many enticing avenues for future research. There is considerable scope for extending the investigation of dual formulations to other penalty functions and other fitting criteria. It would also be valuable to explore a functional formulation of the duality relationship. The extensive literature on covering numbers and Kolmogorov entropy for functions of bounded variation can be deployed to study consistency and rates of convergence. And inevitably there will be questions about automatic λ selection. We hope to be able to address some of these issues in subsequent work.

References

ANDERSEN, E., AND Y. YE (1998): "A computational study of the homogeneous algorithm for large-scale convex optimization," *Computational Optimization and Applications*, 10, 243–269.

DAVIES, P. L., AND A. KOVAC (2001): "Local extremes, runs, strings and multiresolution," *The Annals of Statistics*, 29, 1–65.

——— (2004): "Densities, Spectral Densities and Modality," *The Annals of Statistics*, 32, 1093–1136.

Distribution	TV	TS	K-ucv	K-bcv
MW 1	0.109	0.166	0.089	0.082
MW 2	0.109	0.173	0.099	0.092
MW 3	0.130	0.218	0.191	0.200
MW 4	0.143	0.212	0.199	0.202
MW 5	0.120	0.177	0.150	0.140
MW 6	0.110	0.187	0.105	0.104
MW 7	0.127	0.204	0.120	0.116
MW 8	0.113	0.187	0.116	0.124
MW 9	0.120	0.204	0.118	0.132
MW 10	0.190	0.289	0.190	0.348
MW 11	0.121	0.193	0.118	0.117
MW 12	0.178	0.262	0.182	0.274
MW 13	0.143	0.214	0.146	0.143
MW 14	0.225	0.295	0.222	0.279
MW 15	0.242	0.311	0.224	0.248
MW 16	0.129	0.201	0.140	1.279

TABLE 2. Median Integrated Absolute Error: Sample size, $n = 500$ and number of replications $R = 1000$.

——— (2005): "ftnonpar: Features and Strings for Nonparametric Regression," http://cran.R-project.org.

DONOHO, D. L., I. M. JOHNSTONE, J. C. HOCH, AND A. S. STERN (1992): "Maximum entropy and the nearly black object," *J. R. Stat. Soc. (B)*, 54, 41–67.

DUDÍK, M., S. PHILLIPS, AND R. SCHAPIRE (2004): "Performance Guarantees for Regularized Maximum Entropy Density Estimation," in *Proceedings of the 17th Annual Conference on Computational Learning Theory*, ed. by J. Shawe-Taylor, and Y. Singer.

EFRON, B., AND R. TIBSHIRANI (1996): "Using Specially Designed Exponential Families for Density Estimation," *The Annals of Statistics*, 24, 2431–2461.

EGGERMONT, P., AND V. LARICCIA (2001): *Maximum Penalized Likelihood Estimation*. Springer-Verlag.

GALLANT, A. R., AND D. W. NYCHKA (1987): "Semi-nonparametric maximum likelihood estimation," *Econometrica*, 55, 363–390.

GOOD, I. J. (1971): "A nonparametric roughness penalty for probability densities," *Nature*, 229, 29–30.

GOOD, I. J., AND R. A. GASKINS (1971): "Nonparametric roughness penalties for probability densities," *Biometrika*, 58, 255–277.

GU, C. (2002): *Smoothing spline ANOVA models*. Springer-Verlag.

GU, C. (2005): "gss: An R Package for general smoothing splines," R package version 0.9-3, http://cran.R-project.org.

HALL, P., AND M. C. MINNOTTE (2002): "High order data sharpening for density estimation," *Journal of the Royal Statistical Society, B,*, 64(1), 141–157.

Distribution	TV	TS	K-ucv	K-bcv
MW 1	0.0039	0.0074	0.0021	0.0018
MW 2	0.0042	0.0088	0.0028	0.0024
MW 3	0.0096	0.0468	0.0162	0.0280
MW 4	0.0117	0.0293	0.0163	0.0202
MW 5	0.0241	0.0577	0.0220	0.0183
MW 6	0.0037	0.0090	0.0029	0.0027
MW 7	0.0052	0.0121	0.0041	0.0037
MW 8	0.0042	0.0095	0.0041	0.0050
MW 9	0.0042	0.0104	0.0037	0.0043
MW 10	0.0163	0.0393	0.0137	0.0468
MW 11	0.0049	0.0101	0.0045	0.0043
MW 12	0.0118	0.0225	0.0115	0.0223
MW 13	0.0072	0.0136	0.0073	0.0071
MW 14	0.0200	0.0310	0.0174	0.0276
MW 15	0.0226	0.0334	0.0168	0.0231
MW 16	0.0147	0.0349	0.0145	0.5596

TABLE 3. Median Integrated Squared Error: Sample size, $n = 500$ and number of replications $R = 1000$.

HANSEN, M., AND C. KOOPERBERG (2002): "Spline Adaptation in Extended Linear Models," *Statistical Science*, 17, 2–51.

HANSEN, M., C. KOOPERBERG, AND S. SARDY (1998): "Triogram Models," *J. Am. Stat. Assoc.*, 93, 101–119.

JAYNES, E. (1957): "Information theory and statistical mechanics," *Physics Review*, 106, 620–630.

KLONIAS, V. K. (1991): "The influence function of maximum penalized likelihood density estimators," in *Nonparametric Functional Estimation and Related Topics*, ed. by G. Roussas. Kluwer.

KOENKER, R., AND I. MIZERA (2002): "Comment on Hansen and Kooperberg: Spline Adaptation in Extended Linear Models," *Statistical Science*, 17, 30–31.

——— (2004): "Penalized triograms: total variation regularization for bivariate smoothing," *J. R. Stat. Soc. (B)*, 66, 145–163.

KOENKER, R., P. NG, AND S. PORTNOY (1994): "Quantile Smoothing Splines," *Biometrika*, 81, 673–680.

KOOPERBERG, C., AND C. J. STONE (1991): "A Study of Logspline Density Estimation," *Computational Statistics and Data Analysis*, 12, 327–347.

LEONARD, T. (1978): "Density estimation, stochastic processes and prior information," *J. R. Stat. Soc. (B)*, 40, 113–132.

MÄCHLER, M. (2005): "nor1mix: Normal (1-d) Mixture Models Classes and Methods," R package version 1.0-5, http://cran.R-project.org.

MAMMEN, E., AND S. VAN DE GEER (1997): "Locally Adaptive Regression Splines," *Annals of Statistics*, 25, 387–413.

MARRON, J. S., AND M. P. WAND (1992): "Exact mean integrated squared error," *The Annals of Statistics*, 20, 712–736.

MUMFORD, D., AND J. SHAH (1989): "Optimal approximations by piecewise smooth functions and associated variational problems," *Communications on Pure and Applied Mathematics*, 42, 577–684.

NATANSON, I. (1974): *Theory of Functions of a Real Variable*. Ungar.

O'SULLIVAN, F. (1988): "Fast computation of fully automated log-density and log-hazard estimators," *SIAM Journal on Scientific and Statistical Computing*, 9, 363–379.

RAMSAY, J. O., AND B. W. SILVERMAN (2002): *Applied Functional Data Analysis*. Springer, New York.

RUDIN, L., S. OSHER, AND E. FATEMI (1992): "Nonlinear total variation based noise removal algorithms," *Physica D*, 60, 259–268.

RUFIBACH, K., AND L. DÜMBGEN (2004): "Maximum Likelihood Estimation of a Log-Concave Density: Basic Properties and Uniform Consistency," preprint.

SARDY, S., AND P. TSENG (2005): "Estimation of (Non)smooth densities by total variation penalized likelihood," preprint.

SAUNDERS, M. (2004): "PDCO convex optimization software (MATLAB)," http://www.stanford.edu/group/SOL/soft-ware.html.

SCOTT, D. W. (1992): *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons.

SILVERMAN, B. W. (1982): "On the estimation of a probability density function by the maximum penalized likelihood method," *The Annals of Statistics*, 10, 795–810.

——— (1984): "Spline smoothing: The equivalent variable kernel method," *The Annals of Statistics*, 12, 898–916.

——— (1986): *Density estimation for statistics and data analysis*. Chapman & Hall.

STONE, C., M. HANSEN, C. KOOPERBERG, AND Y. TRUONG (1997): "Polynomial splines and their tensor products in extended linear modeling," *Annals of Statistics*, 25, 1371–1470.

THOMPSON, J. R., AND R. A. TAPIA (1990): *Nonparametric function estimation, modeling, and simulation*. SIAM.

VENABLES, W. N., AND B. D. RIPLEY (2002): *Modern applied statistics with S*. Springer-Verlag.

WAHBA, G. (1990): *Spline models for observational data*, CBMS-NSF Regional Conference Series in Applied Mathematics. SIAM.