Quadratic Programming and Penalised Regression

Andrew D. A. C. Smith[1]

[1] University of Bristol
Corresponding author: Andrew Smith, University of Bristol, School of Mathematics, University Walk, Clifton, Bristol, United Kingdom BS8 1TW
Running title: Quadratic programming and penalised regression

Abstract

Quadratic programming is a versatile tool for calculating estimates in penalised regression. It can be used to produce estimates based on L_1 roughness penalties, as in total variation denoising. In particular, it can calculate estimates when the roughness penalty is the total variation of a derivative of the estimate. Combining two roughness penalties, the total variation and total variation of the third derivative, results in an estimate with continuous second derivative but controls the number of spurious local extreme values. A multiresolution criterion may be included in a quadratic program to achieve local smoothing without having to specify smoothing parameters.

1   Introduction

The problem of regression remains an important aspect of statistics. It is an appropriate way to make inferences about relationships between variables. Nonparametric regression is necessary if the relationship cannot be explained by a small number of parameters. Even the most straightforward case of nonparametric regression between two continuous variables is not an easy problem, and remains far from being 'solved'. There are many different methods for nonparametric regression and each behaves differently in different situations. Three broad categories are: kernel smoothing, wavelets and penalised regression.

We will focus on nonparametric regression between continuous one-dimensional response and explanatory variables. All of the methods described below can be extended to the case of multi-dimensional explanatory variables. Given a set of response observations $y_1, \ldots, y_n$ corresponding to ordered observations $t_1 < \cdots < t_n$, thought to have been generated by the signal plus noise model

$$y_i = f(t_i) + \varepsilon_i, \quad i = 1, \ldots, n,$$

we wish to find an estimate, $\hat{f}$, of the functional relationship, f, sometimes called the signal function, between the explanatory variable t and response variable y. The noise terms ε_i, for $i = 1, \ldots, n$, are realisations of independent random variables with zero mean.

When approaching data for the first time, it is a good idea to try several different methods but not practical to employ every one. It is more convenient to try a subset of methods with different properties, and one such subset is

penalised regression. This is an umbrella term for methods that add a penalty term, which quantifies the smoothness or sparsity of the estimated signal function, to the residual sum of squares. A penalised regression estimate is found by minimising

$$\frac{1}{2}\sum_{i=1}^{n}(y_i - \hat{f}(t_i))^2 + P(\hat{f}, \lambda)$$

over all possible function $\hat{f} : [t_1, t_n] \to \mathbb{R}$. This is the sum of the residual sum of squares, which represents the distance between observation and estimate, and a roughness penalty, P, which is parameterised by a smoothing parameter, λ. The smoothing parameter controls the tradeoff between distance and roughness. The quality of the estimate is affected by the nature of the roughness penalty and the norm on which it is based. Below we will show that quadratic programming can calculate penalised regression estimates with a variety of roughness penalties, in particular total variation denoising estimates, which use an L_1 penalty.

The quadratic program is a vector optimisation procedure that seeks to minimise a quadratic term subject to a set of linear constraints. A general form for a quadratic program is

$$
\begin{aligned}
\min_{\mathbf{x}} \quad & \tfrac{1}{2}\mathbf{x}^{\mathrm{T}}G\mathbf{x} - \mathbf{d}^{\mathrm{T}}\mathbf{x} \\
\text{subject to} \quad & A^{\mathrm{T}}\mathbf{x} \geq \mathbf{b},
\end{aligned}
\tag{1}
$$

where $\mathbf{d}$ is a vector with the same dimension as $\mathbf{x}$, the Hessian G is a square, symmetric matrix with number of rows equal to the dimension of $\mathbf{x}$, the matrix A has the same number of rows as G, and the vector $\mathbf{b}$, together

with A, encodes constraints on $\mathbf{x}$. This can be solved in finite time by, for example, the numerically stable active set algorithm of Goldfarb and Idnani (1983), which is probably the most frequently implemented algorithm for solving quadratic programs and is therefore used in the examples below.

In many forms of penalised regression the estimate of the signal function is defined by its values at the explanatory observations, so it is sufficient to calculate $f_i = \hat{f}(t_i)$ for $i = 1, \ldots, n$.

2 Total variation denoising with higher-order penalties

Total variation denoising is a type of penalised regression in which the roughness penalty is equal to the total variation of the estimated signal function. The total variation of a function f on the interval $[t_1, t_n]$ is defined as

$$\sup_{N \in \mathbb{N}} \sup_{\boldsymbol{\tau} \in \mathcal{T}(N)} \sum_{i=1}^{N-1} |f(\tau_{i+1}) - f(\tau_i)|$$

where

$$\mathcal{T}(N) = \left\{ \boldsymbol{\tau} = (\tau_1, \ldots, \tau_N) : t_1 = \tau_1 < \cdots < \tau_N = t_n \right\}.$$

When this is used as a roughness penalty the resulting estimate is a piecewise-constant function that changes value only at points in a subset of $t_1, \ldots, t_n$. Therefore the total variation of the estimated signal function may be expressed as $\sum_{i=1}^{n-1} |f_{i+1} - f_i|$. Since it employs the L_1 norm in the roughness penalty, total variation denoising may be thought of as a nonparametric ver-

sion of the lasso estimator (Tibshirani, 1996) and it hence leads to sparse estimates. When approaching new data it is possible that the underlying signal function has a very simple form; sparse estimates can more easily reveal this. Furthermore this penalty also reduces the number of local extreme values in the estimate (Davies and Kovac, 2001). The total variation penalty may be extended to multi-dimensional explanatory variables, for instance in image denoising. This was first put forward by Rudin et al. (1992), and fast algorithms for specific penalties were developed by Chambolle (2004), and Kovac and Smith (2011). This roughness penalty is readily extendable and the resulting estimates may be calculated by quadratic programming.

With a one-dimensional explanatory variable, the total variation penalty can be generalised in two ways: by extending it to higher derivatives of the estimate, and by adding local smoothing parameters. We will construct a roughness penalty, $P_{\text{TV}(p)}$, that is proportional to the total variation of the pth derivative of the estimate. The amount of penalisation can be varied in different locations by using a vector of local smoothing parameters $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_{n-p-1})^{\text{T}}$. The generalised total variation roughness penalty is

$$P_{\text{TV}(p)}(\hat{f}; \boldsymbol{\lambda}) = \| \operatorname{diag}(\boldsymbol{\lambda}) D_{n-p} \Delta_p \mathbf{f} \|_1$$

where $\mathbf{f} = (f_1, \ldots, f_n)^{\mathrm{T}}$,

$$D_n = \begin{pmatrix} -1 & 1 & & & \\ & -1 & 1 & & \\ & & \ddots & \ddots & \\ & & & -1 & 1 \end{pmatrix}$$

is a matrix with $n-1$ rows and n columns, and $\Delta_p = T_p^{-1} D_{n-p+1} \cdots T_1^{-1} D_n$ with $\Delta_0 = I$ and $T_p = \mathrm{diag}\left((t_{1+p} - t_1, t_{2+p} - t_2, \ldots, t_n - t_{n-p})^{\mathrm{T}} \right)$.

To include this roughness penalty in a quadratic program, we must introduce a vector $\mathbf{v} = (v_1, \ldots, v_{n-p-1})^{\mathrm{T}}$ of dummy variables. If $\mathbf{v}$ does not appear anywhere else in the quadratic program, then $P_{\mathrm{TV}(p)}$ may be included in the quadratic program as

$$\boldsymbol{\lambda}^{\mathrm{T}}(2\mathbf{v} - D_{n-p}\Delta_p\mathbf{f}) \text{ subject to } \mathbf{v} \geq \mathbf{0} \text{ and } \mathbf{v} - D_{n-p}\Delta_p\mathbf{f} \geq \mathbf{0}.$$

The residual sum of squares can also be generalised by including individual non-negative weights on each observation, contained in the vector $\mathbf{w} = (w_1, \ldots, w_n)^{\mathrm{T}}$. The generalised sum of squares is $\sum_{i=1}^n w_i(y_i - f_i)^2$, which is equivalent to $(\mathbf{y} - \mathbf{f})^{\mathrm{T}} W (\mathbf{y} - \mathbf{f})$, where $\mathbf{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$ and $W = \mathrm{diag}(\mathbf{w})$. This is easily included in a quadratic program.

Therefore the estimate may be calculated with a quadratic program (1), where $\mathbf{x}^{\mathrm{T}} = (\mathbf{f}^{\mathrm{T}}, \mathbf{v}^{\mathrm{T}})$,

$$G = \begin{pmatrix} W & 0 \\ 0 & 0 \end{pmatrix}, \quad \mathbf{d} = \begin{pmatrix} W\mathbf{y} + \Delta_p^{\mathrm{T}} D_{n-p}^{\mathrm{T}} \boldsymbol{\lambda} \\ -2\boldsymbol{\lambda} \end{pmatrix}, \quad A = \begin{pmatrix} -\Delta_p^{\mathrm{T}} D_{n-p}^{\mathrm{T}} & 0 \\ I & I \end{pmatrix}$$

and $\mathbf{b} = \mathbf{0}$. Quadratic program algorithms typically require G to be positive definite, which does not hold in the quadratic program above. We can circumvent this problem by adding a small value to the diagonal entries of G, for instance $1/1000$ times the absolute value of the smallest non-zero element of G. This does not have a discernable effect on the outcome in any of the examples given below. Alternatively, the quadratic program (1) has a dual formulation (Nocedal and Wright, 1999) and in some cases the dual Hessian matrix is positive definite. When $p = 0$ the dual formulation for the quadratic program above is equivalent to the taut string algorithm (Davies and Kovac, 2001), which is a special case of regression on a graph (Kovac and Smith, 2011).

Figure 1 shows a sequence of estimates calculated using the quadratic program above. The example datasets were generated by the Blocks and Doppler functions (Donoho and Johnstone, 1994) with 500 observations in each dataset. The explanatory observations were fixed at $t_i = i/500$, for $i = 1, \ldots, 500$, and Gaussian noise, with zero mean, was added to create noisy response variables. The standard deviation of the noise is 0.5 in the Blocks data and 0.05 in the Doppler data. Since the standard deviation is constant, it is appropriate to use equal weights on each observation. It is known (Mammen and van de Geer, 1997) that an estimate based on the roughness penalty $P_{\mathrm{TV}(p)}$ will be a piecewise polynomial with continuous $(p-1)$th derivative. When $p = 0$ it can be seen that the estimate is piecewise-constant, which is appropriate for the Blocks data as the signal function is piecewise constant. However, for the Doppler data, in which the signal function is smooth, the estimate 'staircases' in an undesirable manner. When

$p \geq 2$ the estimates appear much smoother, and have continuous derivatives, which is much more appropriate for the Doppler data. However, the Blocks estimates exhibit additional, unwanted local extreme values.

3   Combined roughness penalties

3.1   Combining L_1 penalties

The different qualities of different total variation roughness penalties have advantages and disadvantages. When $p = 0$, and the roughness penalty is the total variation, then it reduces the number of additional, unwanted local extreme values. However, the estimate is piecewise-constant and therefore does not have a smooth appearance. When $p = 3$, and the roughness penalty is the total variation of the third derivative, then the estimated signal function is piecewise-cubic and may be considered to be smooth as it has the same degree of differentiability as a smoothing spline estimate. However, this estimate often exhibits additional local extreme values. If we have to choose one roughness penalty then we are forced to choose between these opposing advantages. However, it is possible to combine both advantages by combining two roughness penalties. Including the total variation prevents additional local extremes, provided the associated smoothing parameter is not too small. Additionally including the total variation of the third derivative causes the estimate to have continuous second derivative, again provided the associated smoothing parameter is not too small. The choice of smoothing parameter is discussed in Section 4. A similar approach was used by Aldrin (2004) in
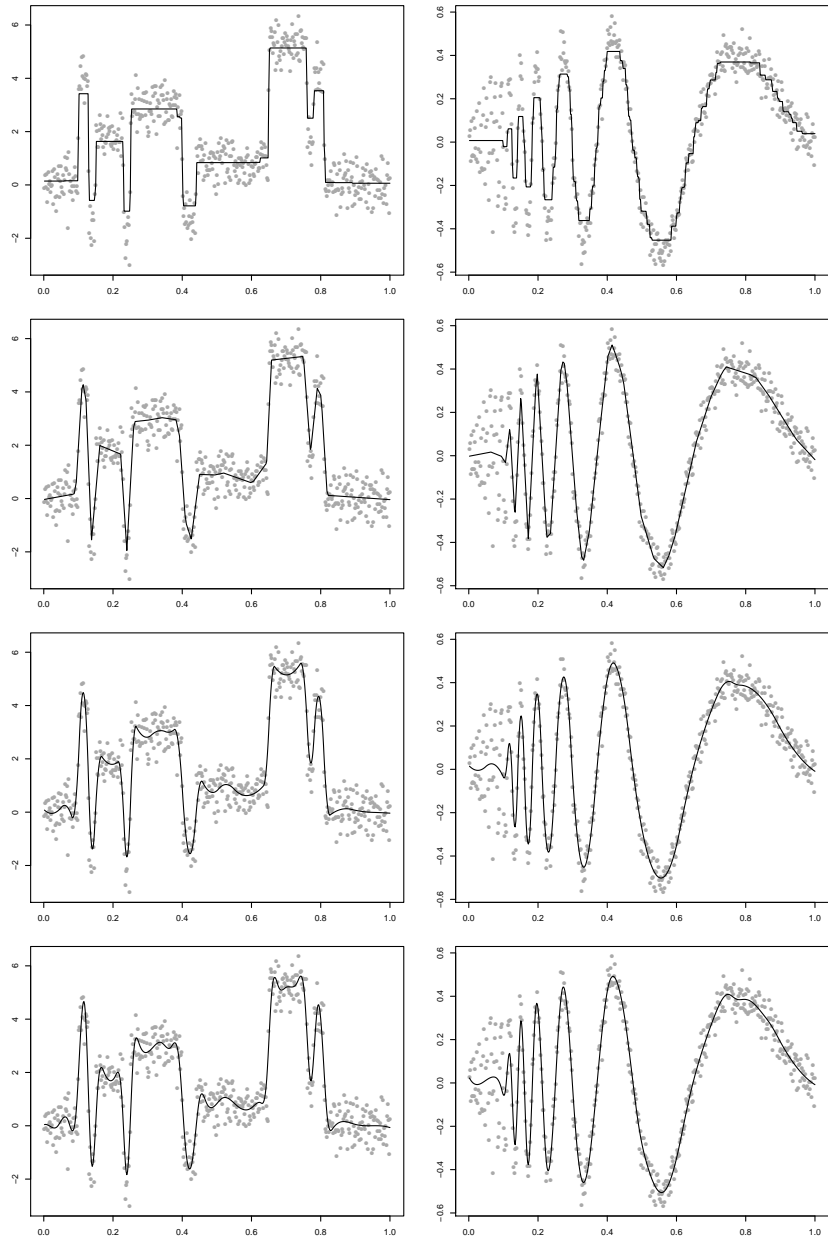
Figure 1: Estimates for the Blocks (left) and Doppler (right) data, based on total variation roughness penalties with (from top to bottom) $p = 1, 2, 3, 4$. Smoothing parameters were chosen for illustrative purposes.

the case of semiparametric regression.

Therefore we seek to minimise

$$\frac{1}{2} \sum_{i=1}^{n} w_i(y_i - f_i)^2 + P_{\text{TV}(0)}(\hat{f}; \boldsymbol{\lambda}) + P_{\text{TV}(3)}(\hat{f}; \boldsymbol{\lambda}').$$

There are separate vectors of smoothing parameters: $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_{n-1})^{\text{T}}$ for $P_{\text{TV}(0)}$ and $\boldsymbol{\lambda}' = (\lambda'_1, \ldots, \lambda'_{n-4})^{\text{T}}$ for $P_{\text{TV}(3)}$.

As before, this minimisation problem can be written as a constrained optimisation problem and hence as a quadratic program. In order to combine two roughness penalties we require two vectors of dummy variables. The first, $\mathbf{v}$, has dimension $n-1$ and the second, $\mathbf{v}'$, has dimension $n-4$. The estimate that we seek can be found by minimising the quadratic program (1), where $\mathbf{x}^{\text{T}} = (\mathbf{f}^{\text{T}}, \mathbf{v}^{\text{T}}, \mathbf{v}'^{\text{T}})$,

$$G = \begin{pmatrix} W & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \mathbf{d} = \begin{pmatrix} W\mathbf{y} + D_n^{\text{T}}\boldsymbol{\lambda} + \Delta_3^{\text{T}} D_{n-3}^{\text{T}} \boldsymbol{\lambda}' \\ -2\boldsymbol{\lambda} \\ -2\boldsymbol{\lambda}' \end{pmatrix},$$

$$A = \begin{pmatrix} -D_n^{\text{T}} & 0 & -\Delta_3^{\text{T}} D_{n-3}^{\text{T}} & 0 \\ I & 0 & I & 0 \\ 0 & I & 0 & I \end{pmatrix}$$

and $\mathbf{b} = \mathbf{0}$.

Examples of this quadratic program, using the Blocks and Doppler datasets seen in Section 2, are given in Figure 2. The practical outcome is as good as the theory suggests: both estimates are smooth, due to the inclusion of
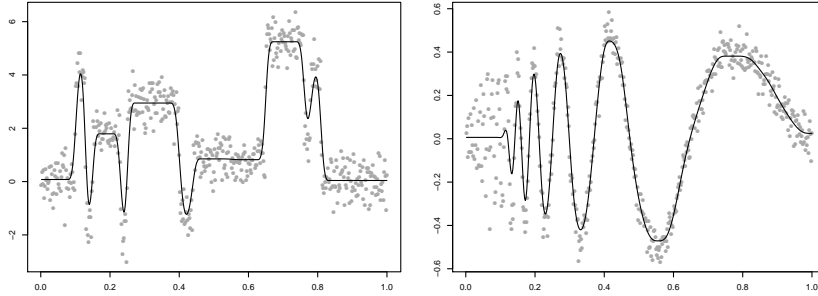
Figure 2: Estimates based on combined roughness penalties for the Blocks (left) and Doppler (right) data. Smoothing parameters were chosen for illustrative purposes.

$P_{\text{TV}(3)}$ as a roughness penalty, but neither exhibits additional, spurious local extreme values, due to the inclusion of $P_{\text{TV}(0)}$ as a roughness penalty.

Also of note is the estimator obtained by combining the residual sum of squares with two roughness penalties $P_{\text{TV}(p)}$ and $P_{\text{TV}(p+1)}$. This is equivalent to a nonparametric version of the fused lasso (Tibshirani et al., 2005).

3.2 Involving L_2 penalties

The estimates above use an L_1 penalty for roughness, as a nonparametric version of the lasso. We can also involve an L_2 penalty as well, expressing it, with local smoothing parameters $\boldsymbol{\lambda}'' = (\lambda_1'', \ldots, \lambda_{n-p-1}'')^{\text{T}}$, as

$$P_{\text{RR}(p)}(\hat{f}, \boldsymbol{\lambda}'') = \|\operatorname{diag}(\boldsymbol{\lambda}'') D_{n-p} \Delta_p \mathbf{f}\|_2^2.$$

Using this roughness penalty alone is a nonparametric version of ridge regression. The minimiser of $\sum_{i=1}^{n} w_i (y_i - f_i)^2 + P_{\text{RR}(p)}(\hat{f}, \boldsymbol{\lambda}'')$ may be found by calculating $(W + 2\Delta_p^{\text{T}} D_{n-p}^{\text{T}} \operatorname{diag}(\boldsymbol{\lambda}'')^2 D_{n-p} \Delta_p)^{-1} W \mathbf{y}$.

11

Alternatively, $P_{\mathrm{RR}(p)}(\hat{f}, \boldsymbol{\lambda}'')$ can be easily added included in the quadratic program of Section 2 by replacing W with $W + 2\Delta_p^{\mathrm{T}} D_{n-p}^{\mathrm{T}} \mathrm{diag}(\boldsymbol{\lambda}'')^2 D_{n-p}\Delta_p$ in the Hessian matrix G. Making this adjustment to the quadratic program in Section 2 will find an estimate that minimises

$$\sum_{i=1}^{n} w_i(y_i - f_i)^2 + P_{\mathrm{TV}(p)}(\hat{f}, \boldsymbol{\lambda}) + P_{\mathrm{RR}(p)}(\hat{f}, \boldsymbol{\lambda}'').$$

As it combines both L_1 and L_2 roughness penalties, this is a nonparametric version of the elastic net (Zou and Hastie, 2005).

4   Multiresolution

It is necessary to make an appropriate choice of smoothing parameter, and there are several techniques for this. Commonly-used techniques designed for the selection of a single, global, smoothing parameter are computationally intensive when generalised to local smoothing parameters. The multiresolution criterion (Davies and Kovac, 2001) can be used to aid the choice of global or local smoothing parameters, without increased complexity for local smoothing. It acts as a test on the residuals $y_i - f_i$ and judges whether they contain any systematic deviations from zero. The residuals are estimates of the noise terms, ε_i, therefore if the residuals are considered to behave like noise then the signal plus noise model is valid and the values f_i are appropriate estimates of $f(t_i)$. If the residuals are considered to contain some of the signal function then the estimates must be too far from the observations and hence the smoothing parameter must be too large. Multiresolution tests

sums of residuals on intervals at different scales and in different locations, and can therefore discern specific intervals in which the smoothing parameter is too large. Through this approach it is possible to choose local smoothing parameters that lead to appropriate, but locally varying, amounts of smoothing in different parts of the estimate.

When the noise terms ε_i are assumed to be independent realisations of a Gaussian random variable with zero mean and constant variance σ^2, then the multiresolution criterion is

$$\left| \frac{\sum_{i=l}^{m} w_i(y_i - f_i)}{\sqrt{\sum_{i=l}^{m} w_i^2}} \right| \leq \hat{\sigma} \sqrt{2 \log n}, \tag{2}$$

where $l = 2^{J-j}(k-1) + 1$ and $m = \min(2^{J-j}k, n)$, for $j = 0, \ldots, J$, where $J = \lceil \log_2 n \rceil$, and $k = 1, \ldots, \lceil 2^{j-J}n \rceil$. If $\sum_{i=l}^{m} w_i^2 = 0$ then the left hand side of (2) is considered to be 0. This requires an estimate of the noise variance that is independent of the residuals. One such estimate is based on the median absolute deviation (MAD):

$$\hat{\sigma} = \frac{\text{median}\left(|y_{i+1} - y_i| : i = 1, \ldots, n-1 \right)}{\sqrt{2}\Phi^{-1}(3/4)}.$$

Typically, the multiresolution criterion is implemented with an iterative procedure (Davies and Kovac, 2001; Davies and Meise, 2008), however, quadratic programming allows a straightforward implementation that makes it possible to avoid specifying a smoothing parameter but still apply an appropriate amount of smoothing. As the multiresolution criterion (2) is a system of linear inequalities it can be directly incorporated into the quadratic pro-

gram. For $j = 0, \ldots, J$ and $k = 1, \ldots, \lceil 2^{j-J} n \rceil$, we append to the columns of A the vectors $W \mathbf{c}_{j,k}$ and $-W \mathbf{c}_{j,k}$ and append to $\mathbf{b}$ the values

$$\mathbf{c}_{j,k}^{\mathrm{T}} W (\mathbf{y} - \hat{\sigma} \sqrt{2 \log n} \mathbf{w}) \text{ and } \mathbf{c}_{j,k}^{\mathrm{T}} W (-\mathbf{y} - \hat{\sigma} \sqrt{2 \log n} \mathbf{w}),$$

where $\mathbf{c}_{j,k}$ is a vector of appropriate dimension with ith element equal to 1 if $2^{J-j}(k-1) + 1 \leq i \leq \min(2^{J-j}k, n)$ and 0 otherwise. To obtain as smooth a fitted function as possible we provide the quadratic program with a large global smoothing parameter, such that the estimate is a straight line in the absence of multiresolution. In the presence of the linear constraints the estimate will be the smoothest function that satisfies (2).

In the case of combined total variation penalties (Section 3.1), we provide two global smoothing parameters λ and λ', and have to choose the ratio λ/λ', specifying how much we favour a smooth estimate over controlling local extremes. This should be chosen within a range that ensures that the estimate will be a straight line in the absence of multiresolution. The specific value of the ratio may be chosen by cross-validation. Alternatively, a pilot estimate may be found using total variation denoising, with $p = 0$, and built-in multiresolution. The resulting estimate is piecewise-constant but may be assumed to have the correct number of local extreme values. A pragmatic approach to setting the value of λ/λ' is to choose a ratio that gives the same number of local extreme values. In practice the specific value of this ratio does not seem to be important; values in $[0.5, 2]$ give no discernable difference for the examples here.

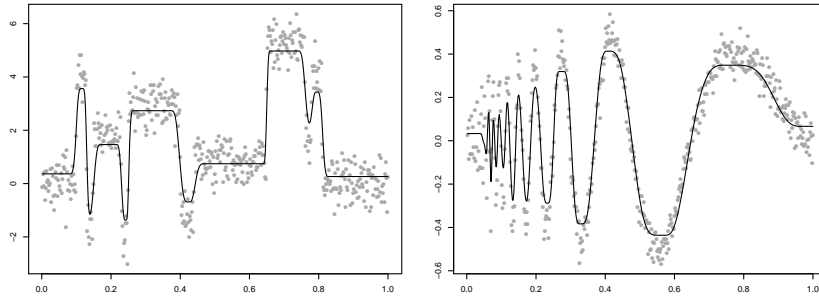Figure 3 shows two examples of estimates calculated using a quadratic

14

Figure 3: Estimates based on combined roughness penalties, with built-in multiresolution based on a system of 1001 intervals, for the Blocks (left) and Doppler (right) data. For each estimate $\lambda = \lambda'$.

program with built-in multiresolution, for the datasets in Figures 1 and 2. Both estimates are smooth but neither exhibits additional, spurious local extreme values. Furthermore the estimates were calculated without any outside specification of the smoothing parameters. It is possible to see the effects of local smoothing in the Doppler estimate: more smoothing has been applied to the right-hand side of the estimate than to the left-hand side. This is appropriate as the signal function is more slowly-varying on the right-hand side than the left-hand side.

5 Discussion

We have seen that it is possible to find estimates based on L_1 penalties without having to choose smoothing parameters, and that quadratic programming is a versatile tool for calculating such estimates. The most appropriate choice of penalty or penalties depends on the qualitative nature of the signal function. If this can be discerned beforehand then it will inform the choice

of penalty. For instance, the discontinuities in the Blocks data above are evident from a scatterplot, hence it is appropriate to apply total variation as a roughness penalty. In contrast, a scatterplot of the Doppler data shows the signal function is continuous and therefore it is more appropriate to apply a higher-order total variation penalty. Quadratic programming allows these different estimates with different qualities to be calculated within the same algorithmic framework.

If the choice of penalty is not clear beforehand, perhaps due to increased noise in the data, then the combined penalty approach of Section 3.1 may be used: if the signal function is smooth it can be approximated by a smooth estimate, but if it has discontinuities it can be estimated without introducing spurious local extreme values. This should only be used if there is no clear information available about the nature of the signal function, which is likely to be a rare occurence.

Estimates based on L_2 penalties, such as ridge regression and the elastic net, can also be incorporated into the quadratic program. Other penalties, such as L_0, may require a different algorithmic approach. Rippe et al. (2012) discuss $P(\hat{f}, \lambda) = \lambda \sum_{i=1}^{n-1} |f_{i+1} - f_i|^q$, where $q \in [0, 1]$, for estimating segmentation (changes in the ratio of alleles) in genomic applications. When $q = 0$ this estimate may be calculated iteratively without using quadratic programming algorithms.

Unfortunately the versatility of quadratic programming must be traded against its computational complexity. The number of calculations performed during the algorithm of Goldfarb and Idnani (1983), for the quadratic program described above, will be $O(n^3)$. The taut string algorithm (Davies and

Kovac, 2001) can find an estimate in $O(n)$ calculations, but only when the roughness penalty is $P_{\text{TV}(0)}$. It would therefore be beneficial, as further work, to develop a faster algorithm specifically designed for the quadratic programs above.

References

Aldrin, M. (2004). Improved predictions penalizing both slope and curvature in additive models. *Comput. Stat. Data An.* 50:267–284.

Chambolle, A. (2004). An algorithm for total variation minimization and applications. *J. Math. Imaging Vis.* 20:89–97.

Davies, P. L., Kovac, A. (2001). Local extremes, runs, strings and multiresolution. *Ann. Statist.* 29:1–65.

Davies, P. L., Meise, M. (2008) Approximating data with weighted smoothing splines. *J. Nonparametr. Stat.* 20:207–228.

Donoho, D. L., Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* 81:425–455.

Goldfarb, D., Idnani, A. (1983). A numerically stable dual method for solving strictly convex quadratic programs. *Math. Program.* 49:325–340.

Kovac, A., Smith, A. D. A. C. (2011) Nonparametric regression on a graph. *J. Comput. Graph. Stat.* 20:432–447.

Mammen, E., van de Geer, S. (1997). Locally adaptive regression splines. *Ann. Statist.* 25:387–413.

Rippe, R. C. A., Meulman, J. J., Eilers, P. H. C. (2012). Visualization of genomic changes by segmented smoothing using an L_0 penalty. *PLoS ONE* 7:e38230.

Rudin, L. I., Osher, S., Fatemi, E. (1992). Nonlinear total variation based noise removal algorithms. *Physica D* 60:259–268.

Nocedal, J., Wright, S. J. (1999). *Numerical Optimization.* New York: Springer.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Stat. Soc. B* 58:267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., Knight, K. (2005) Sparsity and smoothness via the fused lasso. *J. Roy. Stat. Soc. B* 67:91–108.

Zou, H., Hastie, T. (2005) Regularization and variable selection via the elastic net. *J. Roy. Stat. Soc. B* 67:301–320.