SIAM REVIEW
Vol. 51, No. 2, pp. 339–360

ℓ_1 Trend Filtering*

Seung-Jean Kim†
Kwangmoo Koh†
Stephen Boyd†
Dimitry Gorinevsky†

Abstract. The problem of estimating underlying trends in time series data arises in a variety of disciplines. In this paper we propose a variation on Hodrick–Prescott (H-P) filtering, a widely used method for trend estimation. The proposed ℓ_1 *trend filtering method* substitutes a sum of absolute values (i.e., ℓ_1 norm) for the sum of squares used in H-P filtering to penalize variations in the estimated trend. The ℓ_1 trend filtering method produces trend estimates that are piecewise linear, and therefore it is well suited to analyzing time series with an underlying piecewise linear trend. The kinks, knots, or changes in slope of the estimated trend can be interpreted as abrupt changes or events in the underlying dynamics of the time series. Using specialized interior-point methods, ℓ_1 trend filtering can be carried out with not much more effort than H-P filtering; in particular, the number of arithmetic operations required grows linearly with the number of data points. We describe the method and some of its basic properties and give some illustrative examples. We show how the method is related to ℓ_1 regularization-based methods in sparse signal recovery and feature selection, and we list some extensions of the basic method.

Key words. detrending, ℓ_1 regularization, Hodrick–Prescott filtering, piecewise linear fitting, sparse signal recovery, feature selection, time series analysis, trend estimation

AMS subject classifications. 37M10, 62P99

DOI. 10.1137/070690274

1. Introduction.

1.1. Trend Filtering.

We are given a scalar time series y_t, $t = 1, \ldots, n$, assumed to consist of an underlying slowly varying trend x_t and a more rapidly varying random component z_t. Our goal is to estimate the trend component x_t or, equivalently, estimate the random component $z_t = y_t - x_t$. This can be considered as an optimization problem with two competing objectives: We want x_t to be smooth, and we want z_t (our estimate of the random component, sometimes called the *residual*) to be small. In some contexts, estimating x_t is called *smoothing* or *filtering*.

Trend filtering comes up in several applications and settings including macroeconomics (e.g., [52, 86]), geophysics (e.g., [1, 8, 9]), financial time series analysis (e.g., [97]), social sciences (e.g., [66]), revenue management (e.g., [91]), and biological and medical sciences (e.g., [43, 68]). Many trend filtering methods have been proposed,

*Received by the editors May 2, 2007; accepted for publication (in revised form) May 28, 2008; published electronically May 4, 2009. This work was funded in part by the Precourt Institute on Energy Efficiency, by Army award W911NF-07-1-0029, by NSF award 0529426, by NASA award NNX07AEIIA, by AFOSR award FA9550-06-1-0514, and by AFOSR award FA9550-06-1-0312.
http://www.siam.org/journals/sirev/51-2/69027.html

†Information Systems Laboratory, Electrical Engineering Department, Stanford University, Stanford, CA 94305-9510 (sjkim@stanford.edu, deneb1@stanford.edu, boyd@stanford.edu, gorin@stanford.edu).

including Hodrick–Prescott (H-P) filtering [52, 64], moving average filtering [75], exponential smoothing [70], bandpass filtering [21, 4], smoothing splines [81], de-trending via rational square-wave filters [79], a jump process approach [106], median filtering [101], a linear programming (LP) approach with fixed kink points [72], and wavelet transform analysis [23]. (All these methods except for the jump process approach, the LP approach, and median filtering are linear filtering methods; see [4] for a survey of linear filtering methods in trend estimation.) The most widely used methods are moving average filtering, exponential smoothing, and H-P filtering, which is especially popular in economics and related disciplines due to its application to business cycle theory [52]. The idea behind H-P filtering can be found in several fields and can be traced back at least to work in 1961 by Leser [64] in statistics.

1.2. ℓ_1 Trend Filtering. In this paper we propose ℓ_1 *trend filtering*, a variation on H-P filtering which substitutes a sum of absolute values (i.e., an ℓ_1 norm) for the sum of squares used in H-P filtering to penalize variations in the estimated trend. (The term "filtering" is used in analogy with "H-P filtering." Like H-P filtering, ℓ_1 trend filtering is a batch method for estimating the trend component from the whole history of observations.)

We will see that the proposed ℓ_1 trend filter method shares many properties with the H-P filter and has the same (linear) computational complexity. *The principal difference is that the ℓ_1 trend filter produces trend estimates that are smooth in the sense of being piecewise linear.* The ℓ_1 trend filter is thus well suited to analyzing time series with an underlying piecewise linear trend. The kinks, knots, or changes in slope of the estimated trend can be interpreted as abrupt changes or events in the underlying dynamics of the time series; the ℓ_1 trend filter can be interpreted as detecting or estimating changes in an underlying linear trend. Using specialized interior-point methods, ℓ_1 trend filtering can be carried out with not much more effort than H-P filtering; in particular, the number of arithmetic operations required grows linearly with the number of data points.

1.3. Outline. In the next section we set up our notation and give a brief summary of H-P filtering, listing some properties for later comparison with our proposed ℓ_1 trend filter. The ℓ_1 trend filter is described in section 3 and compared to the H-P filter. We give some illustrative examples in section 4.

In section 5 we give the optimality condition for the underlying optimization problem that defines the ℓ_1 trend filter, and we use it to derive some of the properties given in section 3. We also derive a Lagrange dual problem that is interesting on its own and is also used in a primal-dual interior-point method we describe in section 6. We list a number of extensions of the basic idea in section 7.

2. Hodrick–Prescott Filtering. In H-P filtering, the trend estimate x_t is chosen to minimize the weighted sum objective function

$$(1/2)\sum_{t=1}^{n}(y_t - x_t)^2 + \lambda \sum_{t=2}^{n-1}(x_{t-1} - 2x_t + x_{t+1})^2, \tag{1}$$

where $\lambda \geq 0$ is the regularization parameter used to control the trade-off between smoothness of x_t and the size of the residual $y_t - x_t$. The first term in the objective function measures the size of the residual; the second term measures the smoothness of the estimated trend. The argument appearing in the second term, $x_{t-1} - 2x_t + x_{t+1}$, is the second difference of the time series at time t; it is zero when and only when the

three points x_{t-1}, x_t, x_{t+1} are on a line. The second term in the objective is zero if and only if x_t is affine, i.e., has the form $x_t = \alpha + \beta t$ for some constants α and β. (In other words, the graph of x_t is a straight line.) The weighted sum objective (1) is strictly convex and coercive in x, and so has a unique minimizer, which we denote x^{hp}.

We can write the objective (1) as

$$(1/2)\|y - x\|_2^2 + \lambda \|Dx\|_2^2,$$

where $x = (x_1, \ldots, x_n) \in \mathbf{R}^n$, $y = (y_1, \ldots, y_n) \in \mathbf{R}^n$, $\|u\|_2 = (\sum_i u_i^2)^{1/2}$ is the Euclidean or ℓ_2 norm, and $D \in \mathbf{R}^{(n-2)\times n}$ is the second-order difference matrix

$$D = \begin{bmatrix} 1 & -2 & 1 & & & & \\ & 1 & -2 & 1 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & 1 & -2 & 1 & \\ & & & & 1 & -2 & 1 \end{bmatrix}. \tag{2}$$

(D is Toeplitz with first row $[\ 1\ \ -2\ 1\ 0\ \cdots\ 0\ ]$; entries not shown above are zero.) The H-P trend estimate is

$$x^{\mathrm{hp}} = (I + 2\lambda D^T D)^{-1} y. \tag{3}$$

H-P filtering is supported in several standard software packages for statistical data analysis, e.g., `SAS`, `R`, and `Stata`.

We list some basic properties of H-P filtering, which we refer to later when we compare it to our proposed trend estimation method.

- *Linear computational complexity.* The H-P estimated trend x^{hp} in (3) can be computed in $O(n)$ arithmetic operations, since D is tridiagonal.
- *Linearity.* From (3) we see that the H-P estimated trend x^{hp} is a linear function of the time series data y.
- *Convergence to original data as $\lambda \to 0$.* The relative fitting error satisfies the inequality

$$\frac{\|y - x^{\mathrm{hp}}\|_2}{\|y\|_2} \leq \frac{32\lambda}{1 + 32\lambda}. \tag{4}$$

  This shows that as the regularization parameter λ decreases to zero, x^{hp} converges to the original time series data y.
- *Convergence to best affine fit as $\lambda \to \infty$.* As $\lambda \to \infty$, the H-P estimated trend converges to the best affine (straight-line) fit to the time series data,

$$x^{\mathrm{ba}} = \alpha^{\mathrm{ba}} + \beta^{\mathrm{ba}} t,$$

  with intercept and slope

$$\alpha^{\mathrm{ba}} = \frac{\sum_{t=1}^n t^2 \sum_{t=1}^n y_t - \sum_{t=1}^n t \sum_{t=1}^n t y_t}{n \sum_{t=1}^n t^2 - (\sum_{t=1}^n t)^2},$$
$$\beta^{\mathrm{ba}} = \frac{n \sum_{t=1}^n t y_t - \sum_{t=1}^n t \sum_{t=1}^n y_t}{n \sum_{t=1}^n t^2 - (\sum_{t=1}^n t)^2}.$$

- *Commutability with affine adjustment.* We can change the order of H-P filtering and affine adjustment of the original time series data, without affect: For any α and β, the H-P trend estimate of the time series data $\tilde{y}_t = y_t - \alpha - \beta t$ is $\tilde{x}_t^{\rm hp} = x_t^{\rm hp} - \alpha - \beta t$. (A special case of linear adjustment is *linear detrending*, with $\alpha = \alpha^{\rm ba}$, $\beta = \beta^{\rm ba}$, which corresponds to subtracting the best affine fit from the original data.)
- *Regularization path.* The H-P trend estimate $x^{\rm hp}$ is a smooth function of the regularization parameter λ, as it varies over $[0, \infty)$. As λ decreases to zero, $x^{\rm hp}$ converges to the original data y; as λ increases, $x^{\rm hp}$ becomes smoother, and converges to $x^{\rm ba}$, the best affine fit to the time series data.

We can derive the relative fitting error inequality (4) as follows. From the optimality condition $y - x^{\rm hp} = \lambda D^T D x^{\rm hp}$ we obtain

$$y - x^{\rm hp} = 2\lambda D^T D(I + 2\lambda D^T D)^{-1} y.$$

The spectral norm of D is no more than 4:

$$\|Dx\|_2 = \|x_{1:n-2} - 2x_{2:n-1} + x_{3:n}\|_2 \leq \|x_{1:n-2}\|_2 + 2\|x_{2:n-1}\|_2 + \|x_{3:n}\|_2 \leq 4\|x\|_2,$$

where $x_{i:j} = (x_i, \ldots, x_j)$. The eigenvalues of $D^T D$ lie between 0 and 16, so the eigenvalues of $2\lambda D^T D(I + 2\lambda D^T D)^{-1}$ lie between 0 and $32\lambda/(1 + 32\lambda)$. It follows that

$$\|y - x^{\rm hp}\|_2 \leq (32\lambda/(1 + 32\lambda))\|y\|_2.$$

3. ℓ_1 Trend Filtering. We propose the following variation on H-P filtering, which we call ℓ_1 *trend filtering*. We choose the trend estimate as the minimizer of the weighted sum objective function

$$(1/2)\sum_{t=1}^{n}(y_t - x_t)^2 + \lambda \sum_{t=2}^{n-1} |x_{t-1} - 2x_t + x_{t+1}|, \tag{5}$$

which can be written in matrix form as

$$(1/2)\|y - x\|_2^2 + \lambda \|Dx\|_1,$$

where $\|u\|_1 = \sum_i |u_i|$ denotes the ℓ_1 norm of the vector u. As in H-P filtering, λ is a nonnegative parameter used to control the trade-off between smoothness of x and size of the residual. The weighted sum objective (1) is strictly convex and coercive in x and so has a unique minimizer, which we denote $x^{\rm lt}$. (The superscript "lt" stands for "ℓ_1 trend.")

We list some basic properties of ℓ_1 trend filtering, pointing out similarities and differences with H-P filtering.

- *Linear computational complexity.* There is no analytic formula or expression for $x^{\rm lt}$, analogous to (3). But like $x^{\rm hp}$, $x^{\rm lt}$ can be computed numerically in $O(n)$ arithmetic operations. (We describe an efficient method for computing $x^{\rm lt}$ in section 6. Its worst-case complexity is $O(n^{1.5})$, but practically its computational effort is linear in n.)
- *Nonlinearity.* The ℓ_1 trend estimate $x^{\rm lt}$ is *not* a linear function of the original data y. (In contrast, $x^{\rm hp}$ is a linear function of y.)

- *Convergence to original data as* $\lambda \to 0$. The maximum fitting error satisfies the bound

$$\|y - x^{\mathrm{lt}}\|_\infty \leq 4\lambda, \tag{6}$$

  where $\|u\|_\infty = \max_i |u_i|$ denotes the ℓ_∞ norm of the vector u. (Cf. the analogous bound for H-P trend estimation, given in (4).) This implies that $x^{\mathrm{lt}} \to y$ as $\lambda \to 0$.
- *Finite convergence to best affine fit as* $\lambda \to \infty$. As in H-P filtering, $x^{\mathrm{lt}} \to x^{\mathrm{ba}}$ as $\lambda \to \infty$. For ℓ_1 trend estimation, however, the convergence occurs for a finite value of λ,

$$\lambda_{\max} = \|(DD^T)^{-1}Dy\|_\infty. \tag{7}$$

  For $\lambda \geq \lambda_{\max}$, we have $x^{\mathrm{lt}} = x^{\mathrm{ba}}$. (In contrast, $x^{\mathrm{hp}} \to x^{\mathrm{ba}}$ only in the limit as $\lambda \to \infty$.) This maximum value $\lambda_{\max}$ is readily computed with $O(n)$ arithmetic steps. (The derivation is given in section 5.1.)
- *Commutability with affine adjustment.* As in H-P filtering, we can swap the order of affine adjustment and trend filtering, without affect.
- *Piecewise-linear regularization path.* The ℓ_1 trend estimate x^{lt} is a piecewise-linear function of the regularization parameter λ, as it varies over $[0, \infty)$: There are values $\lambda_1, \ldots, \lambda_k$, with $0 = \lambda_k < \cdots < \lambda_1 = \lambda_{\max}$, for which

$$x^{\mathrm{lt}} = \frac{\lambda_i - \lambda}{\lambda_i - \lambda_{i+1}} x^{(i+1)} + \frac{\lambda - \lambda_{i+1}}{\lambda_i - \lambda_{i+1}} x^{(i)}, \quad \lambda_{i+1} \leq \lambda \leq \lambda_i, \quad i = 1, \ldots, k-1,$$

  where $x^{(i)}$ is x^{lt} with $\lambda = \lambda_i$. (So $x^{(1)} = x^{\mathrm{ba}}$, $x^{(k)} = y$.)
- *Linear extension property.* Let $\tilde{x}^{\mathrm{lt}}$ denote the ℓ_1 trend estimate for $(y_1, \ldots, y_{n+1})$. There is an interval $[l, u]$, with $l < u$, for which

$$\tilde{x}^{\mathrm{lt}} = (x^{\mathrm{lt}}, 2x_n^{\mathrm{lt}} - x_{n-1}^{\mathrm{lt}}),$$

  provided $y_{n+1} \in [u, l]$. In other words, if the new observation is inside an interval, the ℓ_1 trend estimate linearly extends the last affine segment.

3.1. Piecewise Linearity. The basic reason the ℓ_1 trend estimate x^{lt} might be preferred over the H-P trend estimate x^{hp} is that it is *piecewise linear* in t: There are (integer) times $1 = t_1 < t_2 < \cdots < t_{p-1} < t_p = n$ for which

$$x_t^{\mathrm{lt}} = \alpha_k + \beta_k t, \quad t_k \leq t \leq t_{k+1}, \quad k = 1, \ldots, p-1. \tag{8}$$

In other words, over each (integer) interval $[t_i, t_{i+1}]$, x^{lt} is an affine function of t. We can interpret α_k and β_k as the local intercept and slope in the kth interval. These local trend parameters are not independent: they must give consistent values for x^{lt} at the join or kink points, i.e.,

$$\alpha_k + \beta_k t_{k+1} = \alpha_{k+1} + \beta_{k+1} t_{k+1}, \quad k = 1, \ldots, p-1.$$

The points $t_2, \ldots, t_{p-1}$ are called *kink points*. We say that x^{lt} is piecewise linear with $p-2$ kink points. (The kink point t_k can be eliminated if $\alpha_k = \alpha_{k-1}$, so we generally assume that $\alpha_k \neq \alpha_{k-1}$.)

In one extreme case, we have $p = 2$, which corresponds to no kink points. In this case $t_1 = 1$, $t_2 = n$, and $x^{\mathrm{lt}} = x^{\mathrm{ba}}$ is affine. In the other extreme case, there is a kink

at every time point: we have $t_i = i$, $i = 1, \ldots, p = n$. In this case the piecewise linear form (8) is vacuous; it imposes no constraints on x^{lt}. This corresponds to $\lambda = 0$, and $x^{\mathrm{lt}} = y$.

The kink points correspond to changes in slope of the estimated trend and can be interpreted as abrupt changes or events in the underlying dynamics of the time series. The number of kinks in x^{lt} typically decreases as the regularization parameter increases, but counterexamples show this need not happen.

Piecewise linearity of the trend estimate is not surprising: It is well known when an ℓ_1 norm term is added to an objective to be minimized, or constrained, the solution typically has the argument of the ℓ_1 norm term sparse (i.e., with many zero elements). In this context, we would predict that Dx (the second-order difference of the estimated trend) will have many zero elements, which means that the estimated trend is piecewise linear.

The general idea of ℓ_1 regularization for the purposes of sparse signal recovery or feature selection has been used in geophysics since the early 1970s; see, e.g., [22, 67, 92]. In signal processing, the idea of ℓ_1 regularization comes up in several contexts, including basis pursuit (denoising) [19, 20], image decomposition [31, 88], signal recovery from incomplete measurements [17, 16, 26, 27, 96], sensor selection [55], fault identification [108], and wavelet thresholding [28]. In statistics, the idea of ℓ_1 regularization is used in the well-known Lasso algorithm [93] for ℓ_1-regularized linear regression, its extensions such as the fused Lasso [94], the elastic net [107], the group Lasso [105], and the monotone Lasso [51], and ℓ_1-regularized logistic regression [61, 62, 77]. The idea of ℓ_1 regularization has been used in other contexts, including portfolio optimization [69], control design [48], computer-aided design of integrated circuits [13], decoding of linear codes [15], and machine learning (sparse principal component analysis [25] and graphical model selection [2, 24, 99]).

We note that ℓ_1 trend filtering is related to segmented regression, a statistical regression method in which the variables are segmented into groups and regression analysis is performed on each segment. Segmented regression arises in a variety of contexts, including abrupt change detection and time series segmentation (especially as a preprocessing step for mining time series databases); the reader is referred to a survey [57] and the references therein. There are two types of time series segmentation. One does not require the fits for two consecutive segments to have consistent values at their join point; see, e.g., [71, 63, 87, 80, 102]. The other requires the fits for two consecutive segments to be consistent at their join point, which is often called joinpoint regression; see, e.g., [32, 35, 36, 58, 89, 104]. We can think of ℓ_1 trend filtering as producing a segmented linear regression, with an affine fit on each segment, and with consistent values at the join points. In ℓ_1 trend filtering, the segmentation and the affine fit on each segment are found by solving one optimization problem.

In time series segmentation, we can use the principle of dynamic programming (DP) to find the best fit that minimizes the fitting error among all functions that consist of k affine segments, with or without the requirement of consistency at the join points. In an early paper [5, 6], Bellman showed how DP can be used for segmented linear regression without the requirement of consistency at the join points. The DP argument can find the best fit in $O(n^2k)$ arithmetic operations [44]. The DP algorithm with the consistency requirement at the join points is, however, far more involved than in the case when it is absent. As a heuristic, ℓ_1 trend filtering produces a segmented linear regression in $O(n)$ arithmetic operations. Another heuristic based on grid search is described in [58], and an implementation, called the Joinpoint Regression Program, is available from http://srab.cancer.gov/joinpoint/.

3.2. ℓ_1 Trend Filtering and Sparse Approximation. To see the connection between ℓ_1 trend filtering and ℓ_1 regularization-based sparse approximation more clearly, we note that the ℓ_1 trend filtering problem is equivalent to the ℓ_1-regularized least squares problem:

$$\text{minimize} \quad (1/2)\|A\theta - y\|_2^2 + \lambda \sum_{i=3}^{n} |\theta_i|, \tag{9}$$

where $\theta = (\theta_1, \ldots, \theta_n) \in \mathbf{R}^n$ is the variable and A is the lower triangular matrix

$$A = \begin{bmatrix} 1 & & & & & \\ 1 & 1 & & & & \\ 1 & 2 & 1 & & & \\ 1 & 3 & 2 & \ddots & & \\ \vdots & \vdots & \vdots & \ddots & 1 & \\ 1 & n-1 & n-2 & \cdots & 2 & 1 \end{bmatrix} \in \mathbf{R}^{n \times n}.$$

The solution θ^{lt} to this problem and the ℓ_1 trend estimate are related by

$$x^{\mathrm{lt}} = A\theta^{\mathrm{lt}}. \tag{10}$$

We can give a simple interpretation of the coefficients: θ_1^{lt} is the offset ($\theta_1^{\mathrm{lt}} = x_1^{\mathrm{lt}}$), θ_2^{lt} is the (right) first-order difference at x_1^{lt} ($\theta_2^{\mathrm{lt}} = x_2^{\mathrm{lt}} - x_1^{\mathrm{lt}}$), and for $t \geq 3$, θ_t^{lt} is the second-order difference of x at $t-1$ ($\theta_t^{\mathrm{lt}} = (Dx^{\mathrm{lt}})_{t-2}$). This interpretation shows again the equivalence between the ℓ_1 trend filtering problem and the ℓ_1-regularized least squares problem (9). (This interpretation also shows that ℓ_1 trend filtering is a special type of basis pursuit denoising [20] and is related to multivariate adaptive regression splines (MARS) [37, 49, 50] that use truncated linear functions as basis functions.)

From a standard result in ℓ_1-regularized least squares [30, 83], the solution θ^{lt} to (9) is a piecewise-linear function of the regularization parameter λ, as it varies over $[0, \infty)$. From (10), we can see that the regularization path of ℓ_1 trend filtering is piecewise linear.

4. Illustrative Examples. Our first example uses synthetic data, generated as

$$y_t = x_t + z_t, \quad t = 1, \ldots, n, \qquad x_{t+1} = x_t + v_t, \quad t = 1, \ldots, n-1, \tag{11}$$

with initial condition $x_1 = 0$. Here x_t is the "true" underlying trend, z_t is the irregular component or noise, and v_t is the trend slope. The noises z_t are IID $\mathcal{N}(0, \sigma^2)$. The trend slopes v_t are chosen from a simple Markov process (independent of z). With probability p, we have $v_{t+1} = v_t$, i.e., no slope change in the underlying trend. (Thus, the mean time between slope changes is $1/(1-p)$.) With probability $1-p$, we choose v_{t+1} from a uniform distribution on $[-b, b]$. We choose the initial slope v_1 from a uniform distribution on $[-b, b]$. The change in x_t between two successive changes in slope is given by the product of two independent random variables: the time between changes (which is geometrically distributed with mean $1/(1-p)$) and the slope (which is uniform over $[-b, b]$). It has zero mean and variance $(1+p)(1-p)^{-2}b^2/3$. The standard deviation of the change in x_t between successive changes in slope is thus $\sqrt{(1+p)/3}(b/(1-p))$.

For our example, we use the parameter values

$$n = 1000, \qquad p = 0.99, \qquad \sigma = 20, \qquad b = 0.5.$$

Thus, the mean time between slope changes is 100, and the standard deviation of the change in x_t between slope changes is 40.7. The particular sample we generated had 8 changes in slope.

The ℓ_1 trend estimates were computed using two solvers: `cvx` [42], a MATLAB-based modeling system for convex optimization (which calls `SDPT3` [95] or `SeDuMi` [90], a MATLAB-based solver for convex problems), and a C implementation of the specialized primal-dual interior-point method described in section 6. The run times on a 3GHz Pentium IV were around a few seconds and 0.01 seconds, respectively.

The results are shown in Figure 1. The top left plot shows the true trend x_t, and the top right plot shows the noise corrupted time series y_t. In the middle left plot, we show x^{lt} for $\lambda = 35000$, which results in 4 kink points in the estimated trend. The middle right plot shows the H-P trend estimate with λ adjusted to give the same fitting error as x^{lt}, i.e., $\|y - x^{\mathrm{lt}}\|_2 = \|y - x^{\mathrm{hp}}\|_2$. Even though x^{lt} is not a particularly good estimate of x_t, it has identified some of the slope change points fairly well. The bottom left plot shows x^{lt} for $\lambda = 5000$, which yields 7 kink points in x^{lt}. The bottom right plot shows x^{hp}, with the same fitting error. In this case the estimate of the underlying trend is quite good. Note that the trend estimation error for x^{lt} is better than x^{hp}, especially around the kink points.

Our next example uses real data, 2000 consecutive daily closing values of the S&P 500 Index, from March 25, 1999, to March 9, 2007, after logarithmic transform. The data are shown in the top plot of Figure 2. In the middle plot, we show x^{lt} for $\lambda = 100$, which results in 8 kink points in the estimated trend. The bottom plot shows the H-P trend estimate with the same fitting error.

In this example (in contrast to the previous one) we cannot say that the ℓ_1 trend estimate is better than the H-P trend estimate. Each of the two trend estimates is a smoothed version of the original data; by construction, they have the same ℓ_2 fitting error. If for some reason you believe that the (log of the) S&P 500 Index is driven by an underlying trend that is piecewise linear, you might prefer the ℓ_1 trend estimate over the H-P trend estimate.

5. Optimality Condition and Dual Problem.

5.1. Optimality Condition. The objective function (5) of the ℓ_1 trend filtering problem is convex but not differentiable, so we use a first-order optimality condition based on subdifferential calculus. We obtain the following necessary and sufficient condition for x to minimize (5): there exists $\nu \in \mathbf{R}^n$ such that

$$y - x = D^T\nu, \qquad \nu_t \in \begin{cases} \{+\lambda\}, & (Dx)_t > 0, \\ \{-\lambda\}, & (Dx)_t < 0, \\ [-\lambda, \lambda], & (Dx)_t = 0, \end{cases} \qquad t = 1, \ldots, n-2. \tag{12}$$

(Here, we use the chain rule for subdifferentials: If f is convex, then the subdifferential of $h(x) = f(Ax + b)$ is given by $\partial h(x) = A^T \partial f(Ax + b)$. See, e.g., [7, Prop. B.24] or [10, Chap. 2] for more on subdifferential calculus.) Since DD^T is invertible, the optimality condition (12) can be written as

$$\left((DD^T)^{-1}D(y - x)\right)_t \in \begin{cases} \{+\lambda\}, & (Dx)_t > 0, \\ \{-\lambda\}, & (Dx)_t < 0, \\ [-\lambda,\, \lambda], & (Dx)_t = 0, \end{cases} \qquad t = 1, \ldots, n-2.$$

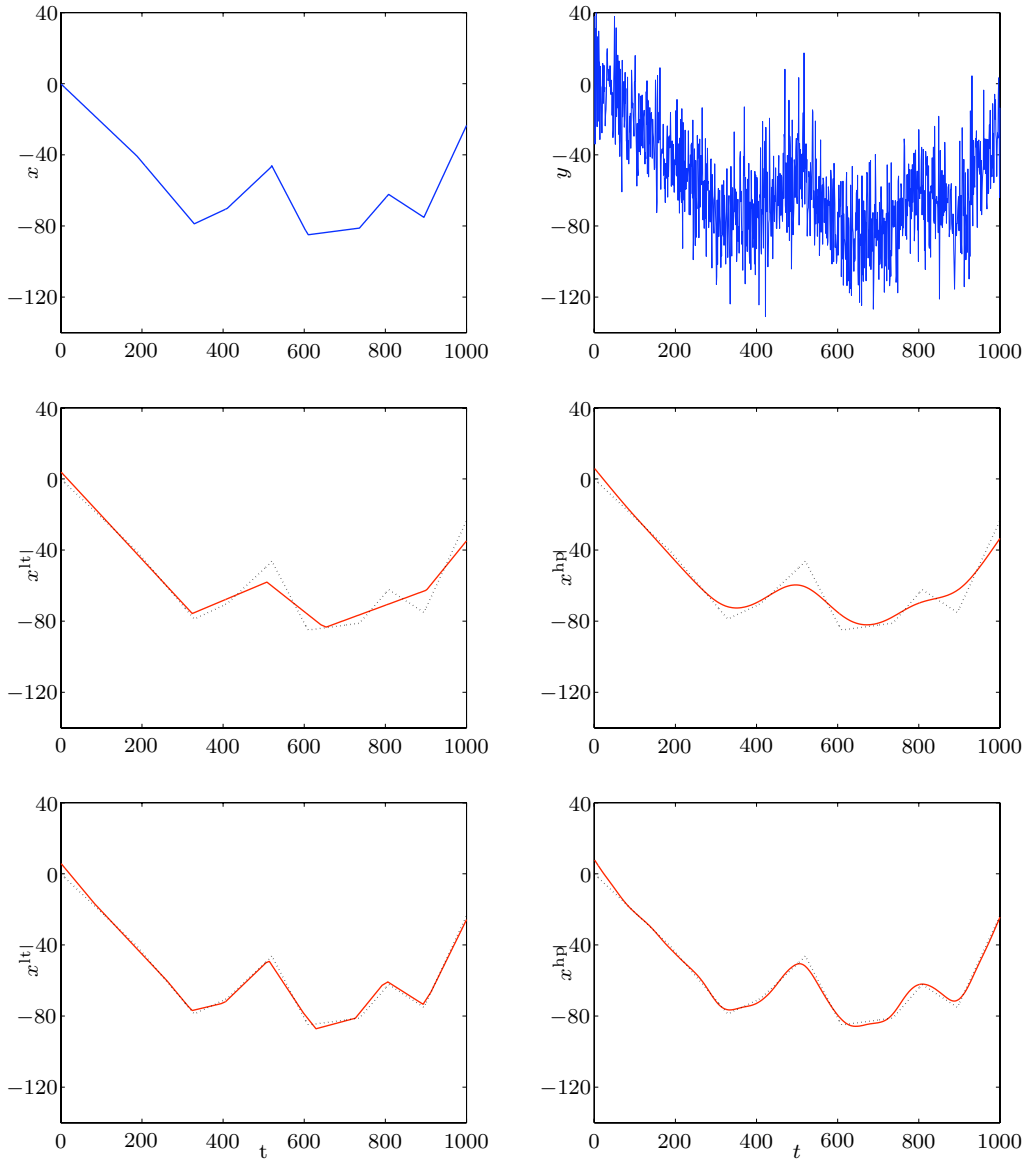

Fig. 1 *Trend estimation on synthetic data. Top left: The true trend x_t. Top right: Observed time series data y_t. Middle left: ℓ_1 trend estimate x^{lt} with four total kinks ($\lambda = 35000$). Middle right: H-P trend estimate x^{hp} with same fitting error. Bottom left: x^{lt} with seven total kinks ($\lambda = 5000$). Bottom right: H-P trend estimate x^{hp} with same fitting error.*

The maximum fitting error bound in (6) follows from the optimality condition above. For any $\nu \in \mathbf{R}^{n-2}$ with $\nu_t \in [-\lambda, \lambda]$,

$$-4\lambda \leq (D^T\nu)_t \leq 4\lambda, \qquad t = 1, \ldots, n.$$

It follows from (12) that the minimizer x of (5) satisfies

$$-4\lambda \leq x_t - y_t \leq 4\lambda, \qquad t = 1, \ldots, n.$$

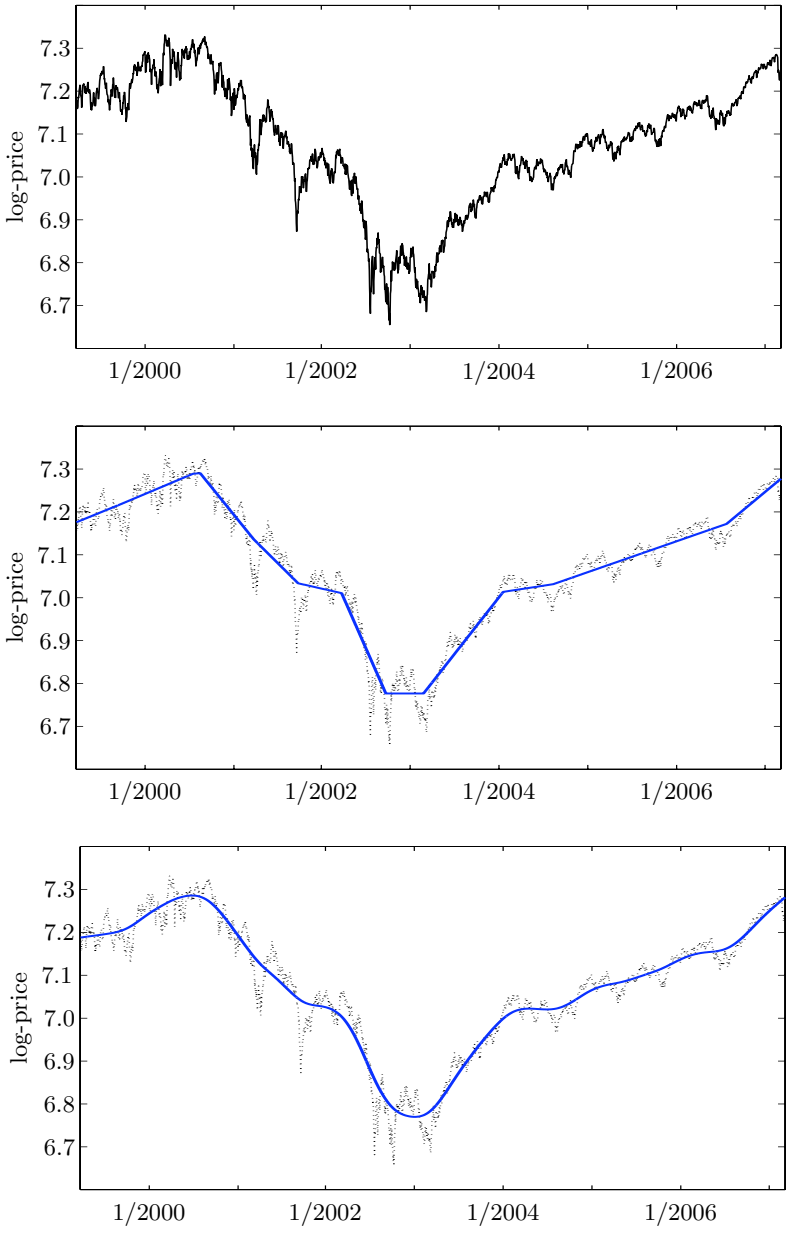

Fig. 2 *Trend estimation results for the S&P* 500 *Index for the period of March* 25, 1999, *to March* 9, 2007. *Top: Original data. Middle:* ℓ_1 *trend estimate* x^{lt} *for* $\lambda = 100$. *Bottom: H-P trend estimate* x^{hp} *with same fitting error.*

We can now derive the formula (7) for $\lambda_{\max}$. Since x^{ba} is affine, $Dx^{\mathrm{ba}} = 0$, so the condition that x^{ba} is optimal is that $\left((DD^T)^{-1}D(y - x^{\mathrm{ba}})\right)_t \in [-\lambda, \lambda]$ for $t = 1, \ldots, n-2$, i.e.,

$$\|(DD^T)^{-1}D(y - x^{\mathrm{ba}})\|_\infty = \|(DD^T)^{-1}Dy\|_\infty \leq \lambda.$$

We can use the optimality condition (12) to see whether the linear extension property holds for a new observation y_{n+1}. From the optimality condition (12), we can see that if y_{n+1} satisfies

$$\left\| (DD^T)^{-1} D \left(\begin{bmatrix} y \\ y_{n+1} \end{bmatrix} - \begin{bmatrix} x^{\mathrm{lt}} \\ 2x_n^{\mathrm{lt}} - x_{n-1}^{\mathrm{lt}} \end{bmatrix} \right) \right\|_\infty \leq \lambda,$$

where $D \in \mathbf{R}^{(n-1)\times(n+1)}$ is the second-order difference matrix on $\mathbf{R}^{n+1}$, then the ℓ_1 trend estimate for (y, y_{n+1}) is given by $(x^{\mathrm{lt}}, 2x_n^{\mathrm{lt}} - x_{n-1}^{\mathrm{lt}}) \in \mathbf{R}^{n+1}$. From this inequality, we can easily find the bounds l and u such that if $l \leq y_{n+1} \leq u$, then the linear extension property holds.

5.2. Dual Problem. To derive a Lagrange dual of the primal problem of minimizing (5), we first introduce a new variable $z \in \mathbf{R}^{n-2}$, as well as a new equality constraint $z = Dx$, to obtain the equivalent formulation

$$\begin{array}{ll} \text{minimize} & (1/2)\|y - x\|_2^2 + \lambda\|z\|_1 \\ \text{subject to} & z = Dx. \end{array}$$

Associating a dual variable $\nu \in \mathbf{R}^{n-2}$ with the equality constraint, the Lagrangian is

$$L(x, z, \nu) = (1/2)\|y - x\|_2^2 + \lambda\|z\|_1 + \nu^T(Dx - z).$$

The dual function is

$$\inf_{x,z} L(x, z, \nu) = \begin{cases} -(1/2)\nu^T DD^T\nu + y^T D^T\nu, & -\lambda\mathbf{1} \leq \nu \leq \lambda\mathbf{1}, \\ -\infty & \text{otherwise.} \end{cases}$$

The dual problem is

$$\begin{array}{ll} \text{minimize} & g(\nu) = (1/2)\nu^T DD^T\nu - y^T D^T\nu \\ \text{subject to} & -\lambda\mathbf{1} \leq \nu \leq \lambda\mathbf{1}. \end{array} \tag{13}$$

(Here $a \leq b$ means $a_i \leq b_i$ for all i.) The dual problem (13) is a (convex) quadratic program (QP) with variable $\nu \in \mathbf{R}^{n-2}$. We say that $\nu \in \mathbf{R}^{n-2}$ is *strictly dual feasible* if it satisfies $-\lambda\mathbf{1} < \nu < \lambda\mathbf{1}$.

From the solution ν^{lt} of the dual (13), we can recover the ℓ_1 trend estimate via

$$x^{\mathrm{lt}} = y - D^T\nu^{\mathrm{lt}}. \tag{14}$$

6. A Primal-Dual Interior-Point Method. The QP (13) can be solved by standard convex optimization methods, including general interior-point methods [12, 73, 74, 103] and more specialized methods such as path following [76, 30]. These methods can exploit the special structure of the problem, i.e., the bandedness of the quadratic form in the objective, to solve the problem very efficiently. To see how this can be done, we describe a simple primal-dual method in this section. For more detail on these (and related) methods, see, e.g., [12, section 11.7] or [103].

The worst-case number of iterations in primal-dual interior-point methods for the QP (13) is $O(n^{1/2})$ [73]. In practice, primal-dual interior-point methods solve QPs in a number of iterations that is just a few tens, almost independent of the problem size or data. Each iteration is dominated by the cost of computing the search direction, which, if done correctly for the particular QP (13), is $O(n)$. It follows that the overall

complexity is $O(n)$, the same as for solving the H-P filtering problem (but with a larger constant hidden in the $O(n)$ notation).

The search direction is the Newton step for the system of nonlinear equations

$$r_t(\nu, \mu_1, \mu_2) = 0, \tag{15}$$

where $t > 0$ is a parameter and

$$r_t(\nu, \mu_1, \mu_2) = \begin{bmatrix} r_{\text{dual}} \\ r_{\text{cent}} \end{bmatrix} = \begin{bmatrix} \nabla g(\nu) + D(\nu - \lambda \mathbf{1})^T \mu_1 - D(\nu + \lambda \mathbf{1})^T \mu_2 \\ -\mu_1(\nu - \lambda \mathbf{1}) + \mu_2(\nu + \lambda \mathbf{1}) - (1/t)\mathbf{1} \end{bmatrix} \tag{16}$$

is the residual. (The first component is the dual feasibility residual, and the second is the centering residual.) Here $\mu_1, \mu_2 \in \mathbf{R}^{n-2}$ are (positive) dual variables for the inequality constraints in (13), and ν is strictly dual feasible. As $t \to \infty$, $r_t(\nu, \mu_1, \mu_2) = 0$ reduces to the Karush–Kuhn–Tucker (KKT) condition for the QP (13). The basic idea is to take Newton steps for solving the set of nonlinear equations $r_t(\nu, \mu_1, \mu_2) = 0$ for a sequence of increasing values of t.

The Newton step is characterized by

$$r_t(\nu + \Delta\nu, \mu_1 + \Delta\mu_1, \mu_1 + \Delta\mu_1) \approx r_t(\nu, \mu_1, \mu_2) + Dr_t(\nu, \mu_1, \mu_2)(\Delta\nu, \Delta\mu_1, \Delta\mu_2) = 0,$$

where Dr_t is the derivative (Jacobian) of r_t. This can be written as

$$\begin{bmatrix} DD^T & I & -I \\ I & J_1 & 0 \\ -I & 0 & J_2 \end{bmatrix} \begin{bmatrix} \Delta\nu \\ \Delta\mu_1 \\ \Delta\mu_2 \end{bmatrix} = - \begin{bmatrix} DD^T z - Dy + \mu_1 - \mu_2 \\ f_1 + (1/t)\operatorname{diag}(\mu_1)^{-1}\mathbf{1} \\ f_2 + (1/t)\operatorname{diag}(\mu_2)^{-1}\mathbf{1} \end{bmatrix}, \tag{17}$$

where

$$\begin{aligned} f_1 &= \nu - \lambda \mathbf{1} \in \mathbf{R}^{n-2}, \\ f_2 &= -\nu - \lambda \mathbf{1} \in \mathbf{R}^{n-2}, \\ J_i &= \operatorname{diag}(\mu_i)^{-1} \operatorname{diag}(f_i) \in \mathbf{R}^{(n-2)\times(n-2)}. \end{aligned}$$

(Here $\operatorname{diag}(w)$ is the diagonal matrix with diagonal entries w.) By eliminating $(\Delta\mu_1, \Delta\mu_2)$, we obtain the reduced system

$$\left(DD^T - J_1^{-1}J_2^{-1}\right)\Delta\nu = -\left(DD^T\nu - Dy - (1/t)\operatorname{diag}(f_1)^{-1}\mathbf{1} + (1/t)\operatorname{diag}(f_2)^{-1}\mathbf{1}\right).$$

The matrix $DD^T - J_1^{-1}J_2^{-1}$ is banded (with bandwidth 5) so we can solve this reduced system in $O(n)$ arithmetic operations. The other two components of the search step, $\Delta\mu_1$ and $\Delta\mu_2$, can be computed as

$$\begin{aligned} \Delta\mu_1 &= -\left(\mu_1 + (1/t)\operatorname{diag}(f_1)^{-1}\mathbf{1} + J_1^{-1}d\nu\right), \\ \Delta\mu_2 &= -\left(\mu_2 + (1/t)\operatorname{diag}(f_2)^{-1}\mathbf{1} - J_2^{-1}d\nu\right) \end{aligned}$$

in $O(n)$ arithmetic operations (since the matrices J_1 and J_2 are diagonal).

A C implementation of a primal-dual interior-point method for ℓ_1 trend filtering is available online from www.stanford.edu/~boyd/l1_tf. For a typical problem with $n = 10000$ data points, it computes x^{lt} in around one second on a 3GHz Pentium IV. Problems with one million data points require around 100 seconds, consistent with linear computational complexity in n.

7. Extensions and Variations. The basic ℓ_1 trend estimation method described above can be extended in many ways, some of which we describe here. In each case, the computation reduces to solving one or a few convex optimization problems, and so is quite tractable; the interior-point method described above is readily extended to handle these problems.

7.1. Polishing. One standard trick is to use the basic ℓ_1 filtering problem as a method to identify the kink points in the estimated trend. Once the kinks points $\{t_1, \ldots, t_p\}$ are identified, we use a standard least-squares method to fit the data over all piecewise-linear functions with the given kinks points:

$$\begin{array}{ll} \text{minimize} & \displaystyle\sum_{k=1}^{p-1} \sum_{t_k \le t \le t_{k+1}} \|y - \alpha_k - \beta_k t\|_2^2 \\ \text{subject to} & \alpha_k + \beta_k t_{k+1} = \alpha_{k+1} + \beta_{k+1} t_{k+1}, \quad k = 1, \ldots, p-2, \end{array}$$

where the variables are the local trend parameters α_k and β_k. This technique is described (in another context) in, e.g., [12, sect. 6.5].

7.2. Iterative Weighted ℓ_1 Heuristic. The basic ℓ_1 trend filtering method is equivalent to

$$\begin{array}{ll} \text{minimize} & \|Dx\|_1 \\ \text{subject to} & \|y - x\|_2 \le s, \end{array} \tag{18}$$

with an appropriate choice of parameter s. In this formulation, we minimize $\|Dx\|_1$ (our measure of smoothness of the estimated trend) subject to a budget on residual norm. This problem can be considered a heuristic for the problem of finding the piecewise-linear trend with the smallest number of kinks, subject to a budget on residual norm:

$$\begin{array}{ll} \text{minimize} & \text{card}(Dx) \\ \text{subject to} & \|y - x\|_2 \le s, \end{array}$$

where $\text{card}(z)$ is the number of nonzero elements in a vector z. Solving this problem exactly is intractable; all known methods require an exhaustive combinatorial search over all—or at least very many—possible combinations of kink points.

The standard heuristic for solving this problem is to replace $\text{card}(Dx)$ with $\|Dx\|_1$, which gives us our basic ℓ_1 trend filter, i.e., the solution to (18). This basic method can be improved by an iterative method that varies the individual weights on the second-order differences in x_t. We start by solving (18). We then define a weight vector as

$$w_t := 1/(\epsilon + |(Dx)_t|), \quad t = 1, \ldots, n-2,$$

where ϵ is a small positive constant. This assigns the largest weight, $1/\epsilon$, when $(Dx)_t = 0$; it assigns large weight when $|(Dx)_t|$ is small; and it assigns relatively small weight when $|(Dx)_t|$ is larger. We then recompute x_t as the solution of problem

$$\begin{array}{ll} \text{minimize} & \|\,\text{diag}(w)Dx\|_1 \\ \text{subject to} & \|y - x\|_2 \le s. \end{array}$$

We then update the weights as above and repeat.

This iteration typically converges in 10 or fewer steps. It often gives a modest decrease in the number of kink points card(Dx), for the same residual, compared to the basic ℓ_1 trend estimation method. The idea behind this heuristic has been used in portfolio optimization with transaction costs [69], where an interpretation of the heuristic for cardinality minimization is given. The reader is referred to [18] for a more extensive discussion on the iterative heuristic.

7.3. Convex Constraints and Penalty Functions. We can add convex constraints on the estimated trend, or use a more general convex penalty function to measure the residual. In both cases, the resulting trend estimation problem is convex, and therefore tractable. We list a few examples here.

Perhaps the simplest constraints are lower and upper bounds on x_t, or the first or second differences of x_t, as in

$$|x_t| \leq M, \quad t = 1, \ldots, n, \qquad |x_{t+1} - x_t| \leq S, \quad t = 1, \ldots, n-1.$$

Here we impose a magnitude limit M, and a maximum slew rate (or slope) S, on the estimated trend. Another interesting convex constraint that can be imposed on x_t is monotonicity, i.e.,

$$x_1 \leq x_2 \leq \cdots \leq x_{n-1} \leq x_n.$$

Minimizing (5) subject to this monotonicity constraint is an extension of isotonic regression, which has been extensively studied in statistics [3, 82]. (Related work on ℓ_1-regularized isotonic regression, in an engineering context, includes [40, 41].)

We can also replace the square function used to penalize the residual term $y_t - x_t$ with a more general convex function ψ. Thus, we compute our trend estimate x_t as the minimizer of (the convex function)

$$\sum_{t=1}^{n} \psi(y_t - x_t) + \lambda \|Dx\|_1.$$

For example, using $\psi(u) = |u|$, we assign a smaller penalty (compared to $\psi(u) = (1/2)u^2$) to large residuals, but a larger penalty to small residuals. This results in a trend estimation method that is more robust to outliers than the basic ℓ_1 trend method since it allows large occasional errors in the residual. Another example is the *Huber penalty function* used in robust least squares, given by

$$\psi_{\mathrm{hub}}(u) = \begin{cases} u^2, & |u| \leq M, \\ M(2|u| - M), & |u| > M, \end{cases}$$

where $M \geq 0$ is a constant [53]. The use of an asymmetric linear penalty function of the form

$$\psi_\tau(u) = \begin{cases} \tau u, & u > 0, \\ -(1-\tau)u & \text{otherwise}, \end{cases}$$

where τ indicates the quantile of interest, is related to quantile smoothing splines. (The reader is referred to [59] for more on the use of this penalty function in quantile regression and [60] for more on quantile smoothing splines.)

In all of these extensions, the resulting convex problem can be solved with a computational effort that is $O(n)$, since the system of equations that must be solved at each step of an interior-point method is banded.

7.4. Multiple Components. We can easily extend basic ℓ_1 trend filtering to analyze time series data that involve other components, e.g., occasional spikes (outliers), level shifts, seasonal components, cyclic (sinusoidal) components, or other regression components. The problem of decomposing given time series data into multiple components has been a topic of extensive research; see, e.g., [11, 29, 45, 46] and the references therein. Compared with standard decomposition methods, the extensions described here are well suited to the case when the underlying trend, once the other components have been subtracted out, is piecewise linear.

Spikes. Suppose the time series data y has occasional spikes or outliers u in addition to trend and irregular components. Our prior information on the component u is that it is sparse. We can extract the underlying trend and the spike signal, by adding one more regularization term to (5), and minimizing the modified objective

$$(1/2)\|y - x - u\|_2^2 + \lambda\|Dx\|_1 + \rho\|u\|_1,$$

where the variables are x (the trend component) and u (the spike component). Here the parameter $\lambda \geq 0$ is used to control the smoothness (or number of slope changes) of the estimated trend, and $\rho \geq 0$ is used to control the number of spikes.

Level Shifts. Suppose the time series data y has occasional abrupt level shifts. Level shifts can be modeled as a piecewise constant component w. To extract the level shift component w as well as the trend x, we add the scaled total variation of w, $\rho\sum_{t=2}^{n}|w_t - w_{t-1}|$, to the weighted sum (5) and minimize the modified objective

$$(1/2)\|y - x - w\|_2^2 + \lambda\|Dx\|_1 + \rho\sum_{t=2}^{n}|w_t - w_{t-1}|,$$

over $x \in \mathbf{R}^n$ and $w \in \mathbf{R}^n$. Here the parameter $\lambda \geq 0$ is used to control the smoothness of the estimated trend x, and $\rho \geq 0$ is used to control the frequency of level shifts in w.

Periodic Components. Suppose the time series data y has an additive deterministic periodic component s with known period p:

$$s_{t+p} = s_t, \quad t = 1, \ldots, n-p.$$

The periodic component s is called "seasonal" when it models seasonal fluctuations; removing effects of the seasonal component from y in order to better estimate the trend component is called *seasonal adjustment*. (The corresponding decomposition problem has been studied extensively in the literature; see, e.g., [14, 29, 49, 56, 65, 85].)

Seasonal adjustment is readily incorporated in ℓ_1 trend filtering: We simply solve the (convex) problem

$$\begin{array}{ll} \text{minimize} & (1/2)\|y - x - s\|_2^2 + \lambda\|Dx\|_1 \\ \text{subject to} & s_{t+p} = s_t, \quad t = 1, \ldots, n-p, \\ & \displaystyle\sum_{k=1}^{p} s_k = 0, \end{array}$$

where the variables are x (the estimated trend) and s (the estimated seasonal component). The last equality constraint means that the periodic component sums to zero over the period; without this constraint, the decomposition is not unique [34, sect. 6.2.8]. To smooth the periodic component, we can add a penalty term to the

objective, or impose a constraint on the variation of s. As a generalization of this formulation, the problem of jointly estimating multiple periodic components (with different periods) as well as a trend can be cast as a convex problem.

When the periodic component is sinusoidal, i.e., $s_t = a\sin\omega t + b\cos\omega t$, where ω is the known frequency, the decomposition problem simplifies to

$$\text{minimize} \quad (1/2)\sum_{t=1}^{n} \|y_t - x_t - a\sin\omega t - b\cos\omega t\|_2^2 + \lambda\|Dx\|_1,$$

where the variables are $x \in \mathbf{R}^n$ and $a, b \in \mathbf{R}$. (H-P filtering has also been extended to estimate trend and cyclic components; see, e.g., [39, 47].)

Regression Components. Suppose that the time series data y has autoregressive (AR) components in addition to the trend x and the irregular component z:

$$y_t = x_t + a_1 y_{t-1} + \cdots + a_r y_{t-r} + z_t,$$

where a_i are model coefficients. (This model is a special type of multiple structural change time series model [100].) We can estimate the trend component and the AR model coefficients by solving the ℓ_1-regularized least squares problem

$$\text{minimize} \quad (1/2)\sum_{i=1}^{n} (y_t - x_t - a_1 y_{t-1} - \cdots - a_r y_{t-r})^2 + \lambda\|Dx\|_1,$$

where the variables are $x_t \in \mathbf{R}^n$ and $a = (a_1, \ldots, a_r) \in \mathbf{R}^r$. (We assume that $y_{1-r}, \ldots, y_0$ are given.)

7.5. Vector Time Series. The basic ℓ_1 trend estimation method can be generalized to handle vector time series data. Suppose that $y_t \in \mathbf{R}^k$ for $t = 1, \ldots, n$. We can find our trend estimate $x_t \in \mathbf{R}^k$, $t = 1, \ldots, k$, as the minimizer of (the convex function)

$$\sum_{t=1}^{n} \|y_t - x_t\|_2^2 + \lambda \sum_{t=2}^{n-1} \|x_{t-1} - 2x_t + x_{t+1}\|_2,$$

where $\lambda \geq 0$ is the usual parameter. Here we use the sum of the ℓ_2 norms of the second differences as our measure of smoothness. (If we use the sum of ℓ_1 norms, then the individual components of x_t can be estimated separately.) Compared to estimating trends separately in each time series, this formulation couples together changes in the slopes of individual entries at the same time index, so the trend component found tends to show simultaneous trend changes, in all components of x_t, at common kink points. (The idea behind this penalty is used in the group Lasso [105] and in compressed sensing involving complex quantities and related to total variation in two- or higher-dimensional data [17, 84].) The common kink points can be interpreted as common abrupt changes or events in the underlying dynamics of the vector time series.

7.6. Spatial Trend Estimation. Suppose we are given two-dimensional data $y_{i,j}$, on a uniform grid $(i, j) \in \{1, \ldots, m\} \times \{1, \ldots, n\}$, assumed to consist of a relatively slowly varying spatial trend component $x_{i,j}$ and a more rapidly varying component $v_{i,j}$. The values of the trend component at node (i, j) and its 4 horizontally or vertically adjacent nodes are on a linear surface if both the horizontal and vertical second-order differences, $x_{i-1,j} - 2x_{ij} + x_{i+1j}$ and $x_{i,j-1} - 2x_{i,j} + x_{i,j+1}$, are zero.

As in the vector time series case, we minimize a weighted sum of the fitting error $\sum_{i=1}^m \sum_{j=1}^n \|y_{i,j} - x_{i,j}\|^2$ and the penalty

$$\sum_{i=2}^{m-1} \sum_{j=2}^{n-1} [(x_{i-1,j} - 2x_{i,j} + x_{i+1,j})^2 + (x_{i,j-1} - 2x_{i,j} + x_{i,j-1})^2]^{1/2}$$

on slope changes in the horizontal and vertical directions. It is possible to use more sophisticated measures of the smoothness, for example, determined by a 9-point approximation that includes 4 diagonally adjacent nodes.

The resulting trend estimates tend to be piecewise linear; i.e., there are regions over which x_t is affine. The boundaries between regions can be interpreted as curves along which the underlying gradient changes rapidly.

7.7. Continuous-Time Trend Filtering. Suppose that we have noisy measurements (t_i, y_i), $i = 1, \ldots, n$, of a slowly varying continuous function at irregularly spaced t_i (in increasing order). In this section we consider the problem of estimating the underlying continuous trend from the finite number of data points. This problem involves an infinite-dimensional set of functions, unlike the trend filtering problems considered above. (Related segmented regression problems have been studied in [38, 54, 63].)

We first consider a penalized least squares problem of the form

$$\text{minimize} \quad (1/2)\sum_{i=1}^n (y_i - x(t_i))^2 + \lambda \int_{t_1}^{t_n} (\ddot{x}(t))^2 \, dt \tag{19}$$

over the space of all functions on the interval $[t_1, t_n]$ with square integrable second derivative. Here, λ is a parameter used to control the smoothness of the solution. The solution is a cubic spline with knots at t_i, i.e., a piecewise polynomial of degree 3 on $\mathbf{R}$ with continuous first and second derivatives; see, e.g., [33, 50, 98]. H-P filtering can be viewed as an approximate discretization of this continuous function estimation problem, when t_i are regularly spaced: $t_i = t_1 + (i-1)h$ for some $h > 0$. If the second derivative of x at t_i is approximated as

$$\ddot{x}(t_i) \approx \frac{x(t_{i-1}) - 2x(t_i) + x(t_{i+1})}{h}, \quad i = 2, \ldots, n-1,$$

then the objective of the continuous-time problem (19) reduces to the weighted sum objective (1) of H-P filtering with regularization parameter λ/h.

We next turn to the continuous time ℓ_1 trend filtering problem

$$\text{minimize} \quad (1/2)\sum_{i=1}^n (y_i - x(t_i))^2 + \lambda \int_{t_1}^{t_n} |\ddot{x}(t)| \, dt \tag{20}$$

over

$$\mathcal{X} = \left\{ x : [t_1, t_n] \to \mathbf{R} \;\middle|\; x(t) = \theta_0 + \theta_1 t + \int_{t_1}^{t_n} \max(t-s, 0) \, d\mu(s),\ \theta_0, \theta_1 \in \mathbf{R},\ V(\mu) < \infty \right\},$$

where $V(\mu)$ is the total variation of the measure μ on $[t_1, t_n]$. (This function space includes piecewise linear continuous functions with a finite number of knots; see [78].) The difference from (19) is that in the integral term the second derivative is penalized using the absolute value function.

A standard result in interpolation theory [78] is that the solution of the interpolation problem

$$\begin{array}{ll}\text{minimize} & \displaystyle\int_{t_1}^{t_n} |\ddot{x}(t)|\,dt \\ \text{subject to} & x(t_i) = y_i, \qquad i = 1, \ldots, n,\end{array}$$

over $\mathcal{X}$ is continuous piecewise linear with knots at the points t_i. From this, we can see that the solution to the continuous time ℓ_1 trend filtering problem (20) is also piecewise continuous linear with knots at the points t_i; i.e., it is a linear spline. The second derivative of a piecewise linear function x with knots at the points t_i is given by

$$\ddot{x}(t) = \sum_{i=2}^{n-1} \left(\frac{x(t_{i+1}) - x(t_i)}{t_{i+1} - t_i} - \frac{x(t_i) - x(t_{i-1})}{t_i - t_{i-1}} \right) \delta(t - t_i),$$

where δ is the Dirac delta function. (The coefficients are slope changes at the kink points.) The integral of the absolute value of the second derivative is

$$\int_{t_1}^{t_n} |\ddot{x}(t)|\,dt = \sum_{i=2}^{n-1} \left| \frac{x(t_{i+1}) - x(t_i)}{t_{i+1} - t_i} - \frac{x(t_i) - x(t_{i-1})}{t_i - t_{i-1}} \right|.$$

Thus the continuous ℓ_1 filtering problem (20) is equivalent to the (finite-dimensional) convex problem

$$\text{minimize} \quad (1/2)\sum_{i=1}^{n}(y_i - x_i)^2 + \lambda \sum_{i=2}^{n-1} \left| \frac{x_{i+1} - x_i}{t_{i+1} - t_i} - \frac{x_i - x_{i-1}}{t_i - t_{i-1}} \right| \tag{21}$$

with variables $(x_1, \ldots, x_n) \in \mathbf{R}^n$. From the optimal points $(t_i, x_i^\star)$, we can easily recover the solution to the original continuous trend filtering problem: the piecewise-linear function that connects $(t_i, x_i^\star)$,

$$x^\star(t) = \frac{t - t_i}{t_{i+1} - t_i} x_{i+1}^\star + \frac{t_{i+1} - t}{t_{i+1} - t_i} x_i^\star, \qquad t \in (t_i, t_{i+1}),$$

is the optimal continuous trend that minimizes (20). When t_i are regularly spaced, this problem reduces to the basic ℓ_1 trend filtering problem considered in section 3. For the same reason, we can solve (21) (and hence (20)) in $O(n)$ arithmetic operations.

7.8. Segmented Polynomial Regression. Thus far our focus has been on fitting a piecewise-linear function to the given data. We can extend the idea to fitting a piecewise polynomial of degree $k-1$ to the data. Using a weighted ℓ_1 norm of the kth-order difference of x as a penalty term, the extension can be formulated as

$$\text{minimize} \quad (1/2)\sum_{i=1}^{n}(y_i - x_i)^2 + \lambda \|D^{(k,n)} x\|_1. \tag{22}$$

Here $D^{(k,n)} \in \mathbf{R}^{(n-k)\times n}$ is the kth-order difference matrix on $\mathbf{R}^n$, defined recursively as

$$D^{(k,n)} = D^{(1,n-k+1)} D^{(k-1,n)}, \quad k = 2, 3, \ldots,$$

where $D^{(1,p)} \in \mathbf{R}^{(p-1)\times p}$ is the first-order difference matrix on $\mathbf{R}^p$,

$$D^{(1,p)} = \begin{bmatrix} 1 & -1 & & & & \\ & 1 & -1 & & & \\ & & \ddots & \ddots & & \\ & & & 1 & -1 & \\ & & & & 1 & -1 \end{bmatrix}.$$

(Entries not shown above are zero.) This problem can be efficiently solved with a computational effort that is $O(k^2 n)$, since the system of equations that must be solved at each step of an interior-point method is banded with bandwidth linear in k.

The resulting trend estimate x_t tends to be piecewise polynomial of order $k-1$, i.e., there are regions over which x_t is polynomial of order $k-1$. The case of $k = 1$ corresponds to piecewise constant fitting and the case of $k = 2$ corresponds to piecewise-linear fitting.

Acknowledgments. The authors thank Trevor Hastie, Johan Lim, Michael Lustig, Almir Mutapcic, and Robert Tibshirani for helpful comments and suggestions.

REFERENCES

[1] R. Baillie and S. Chung, *Modeling and forecasting from trend stationary long memory models with applications to climatology*, Internat. J. Forecast., 18 (2002), pp. 215–226.
[2] O. Banerjee, L. El Ghaoui, A. d'Aspremont, and G. Natsoulis, *Convex optimization techniques for fitting sparse Gaussian graphical models*, in Proceedings of the 23rd International Conference on Machine Learning, 2006.
[3] R. Barlow, D. Bartholomew, J. Bremner, and H. Brunk, *Statistical Inference Under Order Restrictions: The Theory and Application of Isotonic Regression*, John Wiley, New York, 1972.
[4] M. Baxter and R. King, *Measuring business cycles: Approximate band-pass filters for economic time series*, Rev. Econom. Statist., 81 (1999), pp. 575–593.
[5] R. Bellman, *On the approximation of curves by line segments using dynamic programming*, Commun. ACM, 4 (1961), p. 284.
[6] R. Bellman and R. Roth, *Curve fitting by segmented straight lines*, J. Amer. Statist. Assoc., 64 (1969), pp. 1079–1084.
[7] D. Bertsekas, *Nonlinear Programming*, 2nd ed., Athena Scientific, Nashua, NH, 1999.
[8] P. Bloomfeld, *Trends in global temperature*, Climate Change, 21 (1992), pp. 1–16.
[9] P. Bloomfeld and D. Nychka, *Climate spectra and detecting climate change*, Climate Change, 21 (1992), pp. 275–287.
[10] J. Borwein and A. Lewis, *Convex Analysis and Nonlinear Optimization*, Springer-Verlag, New York, 2000.
[11] G. Box, G. Jenkins, and G. Reinsel, *Time Series Analysis: Forecasting & Control*, 3rd ed., Prentice–Hall, Englewood Cliffs, NJ, 1994.
[12] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Cambridge, UK, 2004.
[13] S. Boyd, L. Vandenberghe, A. E. Gamal, and S. Yun, *Design of robust global power and ground networks*, in Proceedings of ACM/SIGDA International Symposium on Physical Design (ISPD), 2001, pp. 60–65.
[14] J. Burman, *Seasonal adjustment by signal extraction*, J. Roy. Statist. Soc. Ser. A, 143 (1980), pp. 321–337.
[15] E. Candès and T. Tao, *Decoding by linear programming*, IEEE Trans. Inform. Theory, 51 (2005), pp. 4203–4215.
[16] E. Candès, J. Romberg, and T. Tao, *Stable signal recovery from incomplete and inaccurate measurements*, Comm. Pure Appl. Math., 59 (2005), pp. 1207–1223.
[17] E. Candès, J. Romberg, and T. Tao, *Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information*, IEEE Trans. Inform. Theory, 52 (2006), pp. 489–509.

[18] E. Candès, E. Wakin, and S. Boyd, *Enhancing sparsity by reweighted ℓ_1 minimization*, J. Fourier Anal. Appl., 14 (2008), pp. 877–905.
[19] S. Chen and D. Donoho, *Basis pursuit*, in Proceedings of the Twenty-Eighth Asilomar Conference on Signals, Systems and Computers, Vol. 1, 1994, pp. 41–44.
[20] S. S. Chen, D. L. Donoho, and M. A. Saunders, *Atomic decomposition by basis pursuit*, SIAM Rev., 43 (2001), pp. 129–159.
[21] L. Christiano and T. Fitzgerald, *The band-pass filter*, Internat. Econom. Rev., 44 (2003), pp. 435–465.
[22] J. Claerbout and F. Muir, *Robust modeling of erratic data*, Geophys., 38 (1973), pp. 826–844.
[23] P. Craigmile, P. Guttorp, and D. Percival, *Trend assessment in a long memory dependence model using the discrete wavelet transform*, Environmetrics, 15 (2004), pp. 313–335.
[24] J. Dahl, L. Vandenberghe, and V. Roychowdhury, *Covariance selection for non-chordal graphs via chordal embedding*, Optim. Methods Softw., 23 (2008), pp. 501–520.
[25] A. d'Aspremont, L. El Ghaoui, M. I. Jordan, and G. R. G. Lanckriet, *A direct formulation for sparse PCA using semidefinite programming*, SIAM Rev., 49 (2007), pp. 434–448.
[26] D. Donoho, *Compressed sensing*, IEEE Trans. Inform. Theory, 52 (2006), pp. 1289–1306.
[27] D. Donoho, M. Elad, and V. Temlyakov, *Stable recovery of sparse overcomplete representations in the presence of noise*, IEEE Trans. Inform. Theory, 52 (2006), pp. 6–18.
[28] D. Donoho, I. Johnstone, G. Kerkyacharian, and D. Picard, *Wavelet shrinkage: Asymptopia?*, J. Roy. Statist. Soc. B., 57 (1995), pp. 301–337.
[29] J. Durbin and S. Koopman, *Time Series Analysis by State Space Methods*, Oxford University Press, Oxford, UK, 2001.
[30] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, *Least angle regression*, Ann. Statist., 32 (2004), pp. 407–499.
[31] M. Elad, J. Starck, D. Donoho, and P. Querre, *Simultaneous cartoon and texture image inpainting using morphological component analysis (MCA)*, Appl. Comput. Harmonic Anal., 19 (2005), pp. 340–358.
[32] J. Ertel and E. Fowlkes, *Some algorithms for linear spline and piecewise multiple linear regression*, J. Amer. Statist. Assoc., 71 (1976), pp. 640–648.
[33] R. Eubank, *Nonparametric Regression and Spline Smoothing*, Marcel Dekker, New York, 1999.
[34] J. Fan and Q. Yao, *Nonlinear Time Series: Nonparametric and Parametric Methods*, Springer-Verlag, New York, 2003.
[35] P. Feder, *The log likelihood ratio in segmented regression*, Ann. Statist., 3 (1975), pp. 84–97.
[36] P. Feder, *On asymptotic distribution theory in segmented regression problems: Identified case*, Ann. Statist., 3 (1975), pp. 49–83.
[37] J. Friedman, *Multivariate adaptive regression splines*, Ann. Statist., 19 (1991), pp. 1–67.
[38] A. Gallant and W. Fuller, *Fitting segmented polynomial regression models whose join points have to be estimated*, J. Amer. Statist. Assoc., 68 (1973), pp. 144–147.
[39] V. Gómez, *The use of Butterworth filters for trend and cycle estimation in economic time series*, J. Bus. Econom. Statist., 19 (2001), pp. 365–373.
[40] D. Gorinevsky, *Monotonic regression filters for trending gradual deterioration faults*, in Proceedings of American Control Conference (ACC), 2004, pp. 5394–5399.
[41] D. Gorinevsky, S.-J. Kim, S. Beard, S. Boyd, and G. Gordon, *Optimal estimation of deterioration from diagnostic image sequence*, IEEE Trans. Signal Process., 57 (2009), pp. 1030–1043.
[42] M. Grant, S. Boyd, and Y. Ye, `cvx`*: A Matlab Software for Disciplined Convex Programming*, 2007. Available online from www.stanford.edu/~boyd/cvx/.
[43] S. Greenland and M. Longnecker, *Methods for trend estimation from summarized dose-response data, with applications to meta-analysis*, Amer. J. Epidemiology, 135 (1992), pp. 1301–1309.
[44] S. Guthery, *Partition regression*, J. Amer. Statist. Assoc., 69 (1974), pp. 945–947.
[45] J. Hamilton, *Time Series Analysis*, Princeton University Press, Princeton, NJ, 1994.
[46] A. Harvey, *Forecasting, Structural Time Series Models and the Kalman Filter*, Cambridge University Press, Cambridge, UK, 1991.
[47] A. Harvey and T. Trimbur, *General model-based filters for extracting cycles and trends in economic time series*, Rev. Econom. Statist., 85 (2003), pp. 244–255.
[48] A. Hassibi, J. How, and S. Boyd, *Low-authority controller design via convex optimization*, in Proceedings of the IEEE Conference on Decision and Control, 1999, pp. 140–145.
[49] T. Hastie and R. Tibshirani, *Generalized Additive Models*, Chapman & Hall/CRC, London, 1990.

[50] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer Ser. Statist., Springer-Verlag, New York, 2001.
[51] T. Hastie, J. Taylor, R. Tibshirani, and G. Walther, *Forward stagewise regression and the monotone lasso*, Electron. J. Statist., 1 (2007), pp. 1–29.
[52] R. Hodrick and E. Prescott, *Postwar U.S. business cycles: An empirical investigation*, J. Money, Credit, and Banking, 29 (1997), pp. 1–16.
[53] P. Huber, *Robust Statistics*, John Wiley, New York, 1981.
[54] D. Hudson, *Fitting segmented curves whose join points have to be estimated*, J. Amer. Statist. Assoc., 61 (1966), pp. 1097–1129.
[55] S. Joshi and S. Boyd, *Sensor selection via convex optimization*, IEEE Trans. Signal Process., 57 (2009), pp. 451–462.
[56] P. Kenny and J. Durbin, *Local trend estimation and seasonal adjustment of economic and social time series*, J. Roy. Statist. Soc. Ser. A, 145 (1982), pp. 1–41.
[57] E. Keogh, S. Chu, D. Hart, and M. Pazzani, *Segmenting time series: A survey and novel approach*, Data Mining in Time Series Databases, (2004), pp. 1–22.
[58] H.-J. Kim, M. Fay, E. Feuer, and D. Midthune, *Permutation tests for joinpoint regression with applications to cancer rates*, Stat. Med., 19 (2000), pp. 335–351.
[59] R. Koenker and G. Bassett, *Regression quantile*, Econometrica, 46 (1978), pp. 33–50.
[60] R. Koenker, P. Ng, and S. Portnoy, *Quantile smoothing splines*, Biometrika, 81 (1994), pp. 673–680.
[61] K. Koh, S.-J. Kim, and S. Boyd, *An interior-point method for large-scale ℓ_1-regularized logistic regression*, J. Machine Learning Res., 8 (2007), pp. 1519–1555.
[62] S. Lee, H. Lee, P. Abeel, and A. Ng, *Efficient l_1-regularized logistic regression*, in Proceedings of the 21st National Conference on Artificial Intelligence (AAAI-06), 2006.
[63] P. Lerman, *Fitting segmented regression models by grid search*, Appl. Statist., 29 (1980), pp. 77–84.
[64] C. Leser, *A simple method of trend construction*, J. R. Stat. Soc. Ser. B Stat. Methodol., 23 (1961), pp. 91–107.
[65] C. Leser, *Estimation of quasi-linear trend and seasonal variation*, J. Amer. Statist. Assoc., 58 (1963), pp. 1033–1043.
[66] S. Levitt, *Understanding why crime fell in the* 1990*s: Four factors that explain the decline and six that do not*, J. Econom. Perspectives, 18 (2004), pp. 163–190.
[67] S. Levy and P. Fullagar, *Reconstruction of a sparse spike train from a portion of its spectrum and application to high-resolution deconvolution*, Geophys., 46 (1981), pp. 1235–1243.
[68] W. Link and J. Sauer, *Estimating equations estimates of trend*, Bird Populations, 2 (1994), pp. 23–32.
[69] M. Lobo, M. Fazel, and S. Boyd, *Portfolio optimization with linear and fixed transaction costs*, Ann. Oper. Res., 152 (2006), pp. 341–365.
[70] R. Lucas, *Two illustrations of the quantity theory of money*, Amer. Econom. Rev., 70 (1980), pp. 1005–14.
[71] V. McGee and W. Carleton, *Piecewise regression*, J. Amer. Statist. Assoc., 65 (1970), pp. 1109–1124.
[72] G. Mosheiov and A. Raveh, *On trend estimation of time-series: A simpler linear programming approach*, J. Oper. Res. Soc., 48 (1997), pp. 90–96.
[73] Y. Nesterov and A. Nemirovskii, *Interior-Point Polynomial Algorithms in Convex Programming*, Stud. Appl. Math. 13, SIAM, Philadelphia, 1994.
[74] J. Nocedal and S. Wright, *Numerical Optimization*, Springer Ser. Oper. Res., Springer-Verlag, New York, 1999.
[75] D. Osborne, *Moving average detrending and the analysis of business cycles*, Oxford Bull. Econom. Statist., 57 (1995), pp. 547–558.
[76] M. Osborne, B. Presnell, and B. Turlach, *A new approach to variable selection in least squares problems*, IMA J. Numer. Anal., 20 (2000), pp. 389–403.
[77] M. Park and T. Hastie, *An ℓ_1 regularization-path algorithm for generalized linear models*, J. R. Stat. Soc. Ser. B Stat. Methodol., 69 (2007), pp. 659–677.
[78] A. Pinkus, *On smoothest interpolants*, SIAM J. Math. Anal., 19 (1988), pp. 1431–1441.
[79] D. Pollock, *Trend estimation and de-trending via rational square-wave filters*, J. Econom., 99 (2000), pp. 317–334.
[80] R. Quandt, *The estimation of the parameter of a linear regression system obeying two separate regimes*, J. Amer. Statist. Assoc., 53 (1958), pp. 873–880.
[81] C. Reinsch, *Smoothing by spline functions*, Numer. Math., 10 (1976), pp. 177–183.
[82] T. Robertson, F. Wright, and R. Dykstra, *Order Restricted Statistical Inference*, John Wiley, New York, 1988.

[83] S. Rosset and J. Zhu, *Piecewise linear regularized solution paths*, Ann. Statist., 35 (2007), pp. 1012–1030.
[84] L. Rudin, S. Osher, and E. Fatemi, *Nonlinear total variation based noise removal algorithms*, Phys. D, 60 (1992), pp. 259–268.
[85] E. Schlicht, *A seasonal adjustment principle and a seasonal adjustment method derived from this principle*, J. Amer. Statist. Assoc., 76 (1981), pp. 374–378.
[86] K. Singleton, *Econometric issues in the analysis of equilibrium business cycle models*, J. Monetary Econom., 21 (1988), pp. 361–386.
[87] A. Smith and D. Cook, *Straight lines with a change-point: A Bayesian analysis of some renal transplant data*, Appl. Statist., 29 (1980), pp. 180–189.
[88] J. Starck, M. Elad, and D. Donoho, *Image decomposition via the combination of sparse representations and a variational approach*, IEEE Trans. Image Process., 14 (2005), pp. 1570–1582.
[89] H. Stone, *Approximation of curves by line segments*, Math. Comp., 15 (1961), pp. 40–47.
[90] J. Sturm, *Using SEDUMI* 1.02*, a Matlab toolbox for optimization over symmetric cones*, Optim. Methods Softw., 11/12 (1999), pp. 625–653.
[91] R. Talluri and G. van Ryzin, *The Theory and Practice of Revenue Management*, Kluwer Academic, Boston, MA, 2004.
[92] H. Taylor, S. Banks, and J. McCoy, *Deconvolution with the l_1 norm*, Geophys., 44 (1979), pp. 39–52.
[93] R. Tibshirani, *Regression shrinkage and selection via the Lasso*, J. Roy. Statist. Soc. Ser. B, 58 (1996), pp. 267–288.
[94] R. Tibshirani, M. Saunders, S. Rosset, and J. Zhu, *Sparsity and smoothness via the fused Lasso*, J. R. Stat. Soc. Ser. B Stat. Methodol., 67 (2005), pp. 91–108.
[95] K. Toh, R. Tütüncü, and M. Todd, *SDPT3 4.0 (beta) (software package)*, 2001. Available online from http://www.math.nus.edu.sg/~mattohkc/sdpt3.html.
[96] J. Tropp, *Just relax: Convex programming methods for identifying sparse signals in noise*, IEEE Trans. Inform. Theory, 53 (2006), pp. 1030–1051.
[97] R. Tsay, *Analysis of Financial Time Series*, 2nd ed., Wiley-Interscience, Hoboken, NJ, 2005.
[98] G. Wahba, *Spline Models for Observational Data*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 59, SIAM, Philadelphia, 1990.
[99] M. Wainwright, P. Ravikumar, and J. Lafferty, *High-dimensional graphical model selection using ℓ_1-regularized logistic regression*, in Advances in Neural Information Processing Systems (NIPS) 19, 2007, pp. 1465–1472.
[100] J. Wang, *A Bayesian time series model of multiple structural changes in level, trend, and variance*, J. Bus. Econom. Statist., 18 (2000), pp. 374–386.
[101] Y. Wen and B. Zeng, *A simple nonlinear filter for economic time series analysis*, Econom. Lett., 64 (1999), pp. 151–160.
[102] K. Worsley, *Testing for a two-phase multiple regression*, Technometrics, 53 (1983), pp. 34–42.
[103] S. Wright, *Primal-Dual Interior-Point Methods*, SIAM, Philadelphia, 1997.
[104] B. Yu, M. Barrett, H.-J. Kim, and E. Feuer, *Estimating joinpoints in continuous time scale for multiple change-point models*, Comput. Statist. Data Anal., 51 (2007), pp. 2420–2427.
[105] M. Yuan and L. Lin, *Model selection and estimation in regression with grouped variables*, J. R. Stat. Soc. Ser. B Stat. Methodol., 68 (2006), pp. 49–67.
[106] S. Zhao and G. Wei, *Jump process for the trend estimation of time series*, Comput. Statist. Data Anal., 42 (2003), pp. 219–241.
[107] H. Zou and T. Hastie, *Regularization and variable selection via the elastic net*, J. Royal Statist. Soc. Ser. B, 67 (2005), pp. 301–320.
[108] A. Zymnis, S. Boyd, and D. Gorinevsky, *Relaxed maximum a posteriori fault identification*, Signal Process., to appear.