ℓ_1 Trend Filtering

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

April 28, 2007

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.*, an 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 is well suited to analyzing time series with an underlying piecewise linear trend. The kinks, 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 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.

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

*Information Systems Laboratory, Electrical Engineering Department, Stanford University, Stanford, CA 94305-9510 USA. Email: `{sjkim,deneb1,boyd,gorin}@stanford.edu`

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.*, [37, 61]), geophysics (*e.g.*, [1, 6, 7]), financial time series analysis (*e.g.*, [68]), social sciences (*e.g.*, [46]), revenue management (*e.g.*, [64]) and biological sciences (*e.g.*, [47]). Many trend filtering methods have been proposed, including Hodrick-Prescott (H-P) filtering [37, 44], moving average filtering [52], exponential smoothing [49], bandpass filtering [18, 4], smoothing splines [57], de-trending via rational square-wave filters [56], a jump process approach [74], median filtering [71], a linear programming (LP) approach with fixed kink points [50], and wavelet transform analysis [19]. (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 since its application to business cycle theory [37]. 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 [44] in statistics.

1.2 ℓ_1 trend filtering

In this paper we propose ℓ_1 *trend filtering*, a variation on Hodrick-Prescott (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. 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, 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 §3, and compared to the H-P filter. We give some illustrative examples in §4.

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

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

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

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

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, that 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 (2) can be computed in $O(n)$ arithmetic operations, since D is tri-diagonal.
- *Linearity.* From (2) 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

$$\frac{\|y - x^{\rm hp}\|_2}{\|y\|_2} \leq \frac{16\lambda}{1+16\lambda}. \tag{3}$$

  This shows that as the regularization parameter λ decreases to zero, $x^{\rm hp}$ 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,

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

  where the intercept and slope are

$$\begin{aligned} \alpha^{\rm ba} &= \frac{2}{n(n-1)} \left((2n-1) \sum_{t=1}^{n} y_t - 3 \sum_{t=1}^{n} t y_t \right), \\ \beta^{\rm ba} &= \frac{6}{n(n-1)} \left(\frac{2}{n+1} \sum_{t=1}^{n} t y_t - \sum_{t=1}^{n} y_t \right). \end{aligned}$$

- *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 $\tilde{x}_t = x_t - \alpha - \beta t$ 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 (3) as follows. From the optimality conditions $y - x^{\rm hp} = \lambda D^T D x^{\rm hp}$ we obtain

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

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 $\lambda D^T D (I + \lambda D^T D)^{-1}$ lie between 0 and $16\lambda/(1+16\lambda)$. It follows that

$$\|y - x^{\rm hp}\|_2 \leq 16\lambda/(1+16\lambda) \|y\|_2.$$

3 ℓ_1 trend filtering

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

$$(1/2)\sum_{t=1}^{n}(y_t - x_t)^2 + \lambda\sum_{t=2}^{n-1}|x_{t-1} - 2x_t + x_{t+1}|, \tag{4}$$

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^{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^{lt}, analogous to (2). But like x^{hp}, x^{lt} can be computed numerically in $O(n)$ arithmetic operations. (We describe an efficient method for computing x^{lt} in §6.)

- *Nonlinearity.* The ℓ_1 trend estimate x^{lt} is *not* a linear function of the original data y. (In contrast, x^{hp} is a linear function of y.)

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

$$\|y - x^{\mathrm{lt}}\|_\infty \le 4\lambda. \tag{5}$$

  (Cf. the analogous bound for H-P trend estimation, given in (3).) 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,

$$\lambda_{\max} = \|(DD^T)^{-1}Dy\|_\infty. \tag{6}$$

  For $\lambda \ge \lambda_{\max}$, 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 §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 $x^{\rm lt}$ 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

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

  where $x^{(i)}$ is $x^{\rm lt}$ with $\lambda = \lambda_i$. (So $x^{(1)} = x^{\rm ba}$, $x^{(k)} = y$.)

3.1 Piecewise linearity

The basic reason the ℓ_1 trend estimate $x^{\rm lt}$ might be preferred over the H-P trend estimate $x^{\rm hp}$ 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

$$x^{\rm lt}_t = \alpha_k + \beta_k t, \quad t_k \le t \le t_{k+1}, \quad k = 1, \ldots, p-1. \tag{7}$$

In other words, over each (integer) interval $[t_i, t_{i+1}]$, $x^{\rm lt}$ 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 $x^{\rm lt}$ at the 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 $x^{\rm lt}$ 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 $\alpha_k \ne \alpha_{k-1}$.)

In one extreme case, we have $p = 2$, which corresponds to no kink points. In this case $t_1 = 1$, $t_2 = n$, and $x^{\rm lt} = x^{\rm ba}$ 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 (7) is vacuous; it imposes no constraints on $x^{\rm lt}$. This corresponds to $\lambda = 0$, and $x^{\rm lt} = y$.

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 $x^{\rm lt}$ 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.

In signal processing, the idea of ℓ_1 regularization comes up in several contexts including basis pursuit (denoising) [16, 17], image decomposition [27, 62], signal recovery from incomplete measurements [14, 13, 22, 23, 67], and wavelet thresholding [24]. In statistics, the idea of ℓ_1 regularization is used in the well-known Lasso algorithm [65] for ℓ_1-regularized linear regression, its extensions such as the fused Lasso [66], the elastic net [75], and the monotone Lasso [36], and ℓ_1-regularized logistic regression [42, 43, 54]. The idea of ℓ_1 regularization has been used in other contexts including portfolio optimization [48], control design [35], computer-aided design of integrated circuits [11], decoding of linear codes [15], and machine learning (sparse principal component analysis [21] and graphical model selection [2, 20, 69]).

3.2 ℓ_1 trend filtering and sparse approximation

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

$$\text{minimize} \quad (1/2)\|A\theta - y\|_2^2 + \lambda \sum_{i=3}^{n} |\theta_i|, \tag{8}$$

where $\theta = (\theta_1, \dots, \theta_n) \in \mathbf{R}^n$ 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

$$x^{\text{lt}} = A\theta^{\text{lt}}. \tag{9}$$

We can give a simple interpretation of the coefficients: θ_1^{lt} is the offset ($\theta_1^{\text{lt}} = x_1^{\text{lt}}$), θ_1^{lt} is the (right) first-order difference at x_1^{lt} ($\theta_2^{\text{lt}} = x_2^{\text{lt}} - x_1^{\text{lt}}$), and for $t \geq 3$, θ_t^{lt} is the second-order difference of x at $t-1$ ($\theta_t^{\text{lt}} = (Dx^{\text{lt}})_{t-2}$). This interpretation shows again the equivalence between the ℓ_1 trend filtering problem and the ℓ_1-regularized least squares problem (8).

From a standard result in ℓ_1-regularized least squares [26, 59], the solution θ^{lt} to (8) is a piecewise-linear function of the regularization parameter λ, as it varies over $[0, \infty)$. From (9), 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

$$y_t = x_t + z_t, \quad t = 1, \dots, n, \qquad x_{t+1} = x_t + v_t, \quad t = 1, \dots, n-1, \tag{10}$$

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` [32], a Matlab-based modeling system for convex optimization (which calls `SeDuMi` [63], a Matlab-based solver for convex problems), and a Matlab implementation of the specialized primal-dual interior-point method described in §6. The run times on 3Ghz Pentium IV were around a few seconds and 0.1 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 $x^{\rm lt}$ 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 $x^{\rm lt}$, *i.e.*, $\|y - x^{\rm lt}\|_2 = \|y - x^{\rm hp}\|_2$. Even though $x^{\rm lt}$ 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 $x^{\rm lt}$ for $\lambda = 5000$, which yields 7 kink points in $x^{\rm lt}$. The bottom right plot shows $x^{\rm hp}$, with the same fitting error. In this case the estimate of the underlying trend is quite good. Note that the trend estimation error for $x^{\rm lt}$ is better than $x^{\rm hp}$, especially around the kink points.

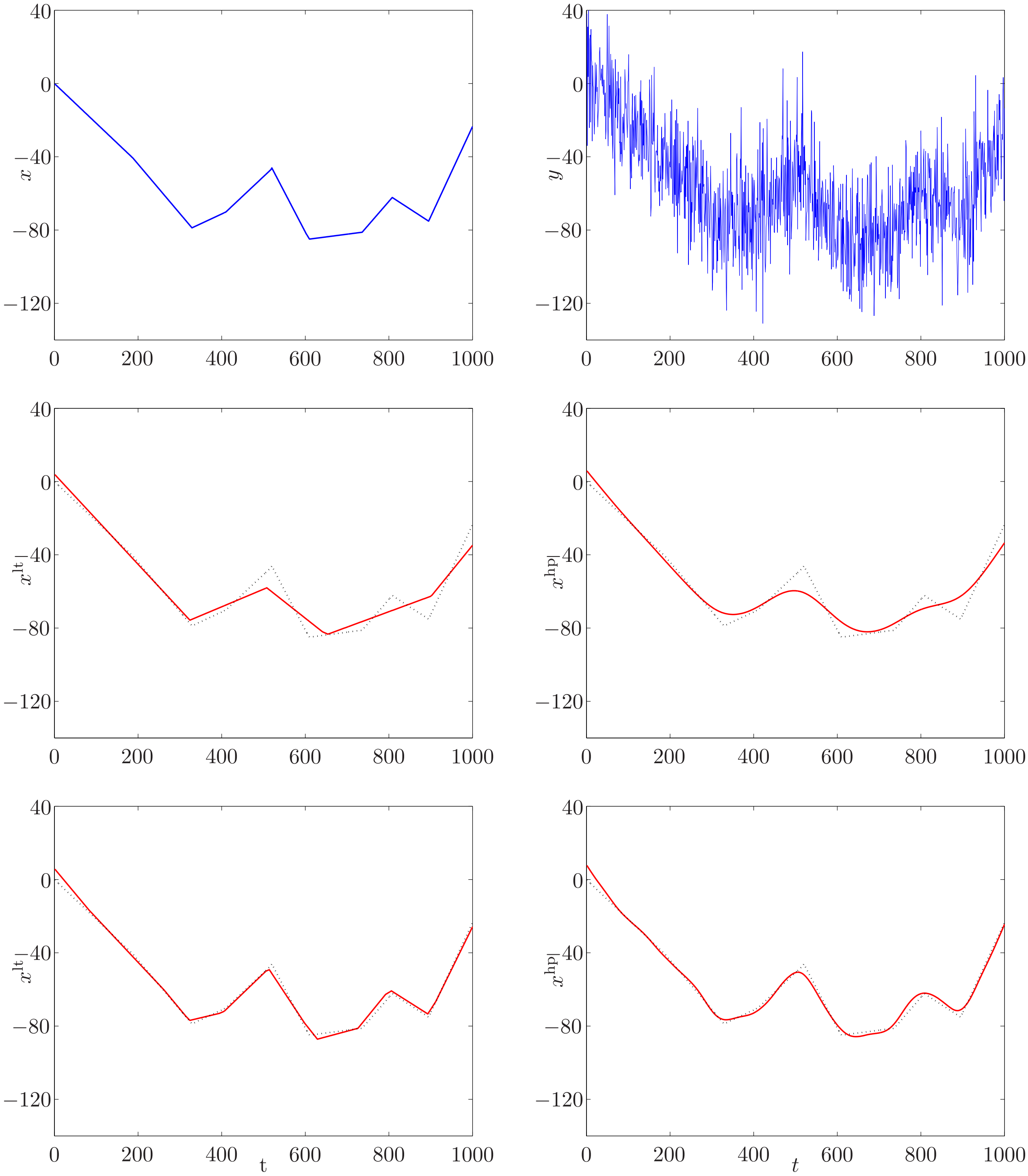

Figure 1: Trend estimation on synthetic data. *Top left.* The true trend x_t. *Top right.* Observed time series data y_t. *Middle left.* ℓ_1 trend estimate x^{lt} with four total kinks ($\lambda = 35000$). *Middle right.* H-P trend estimate x^{hp} with same fitting error. *Bottom left.* x^{lt} with seven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

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.

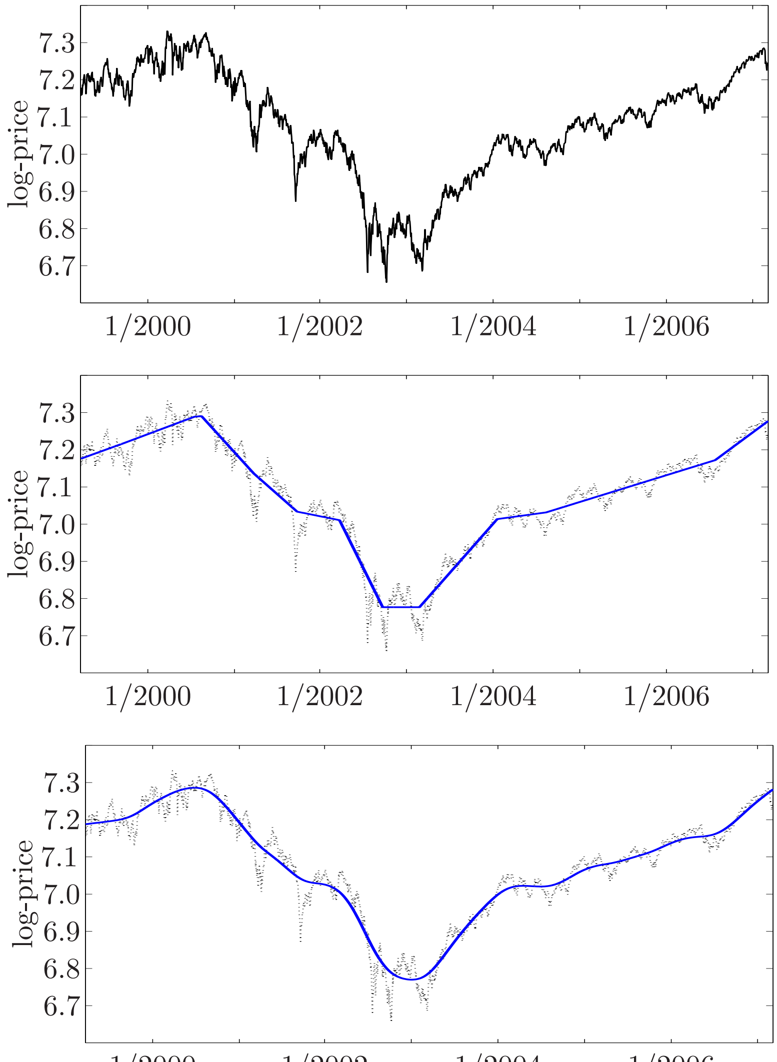

Figure 2: Trend estimation results for the S&P 500 index for the period of March 25, 1999 to March 9, 2007. *Top.* Original data. *Middle.* ℓ_1 trend estimate x^{lt} for $\lambda = 100$. *Bottom.* H-P trend estimate x^{hp} with same fitting error.

5 Optimality condition and dual problem

5.1 Optimality condition

The objective function (4) of ℓ_1 trend filtering 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 (4): there exists $\nu \in \mathbf{R}^n$ such that

$$y - x = D^T\nu, \qquad \nu_t \in \begin{cases} \{\lambda\}, & (Dx)_t > 0, \\ \{-\lambda\}, & (Dx)_t < 0, \\ [-\lambda, \lambda], & (Dx)_t = 0, \end{cases} \qquad t = 1, \ldots, n-2. \tag{11}$$

(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.*, [5, Prop. B.24] or [8, Chap.2] for more on subdifferential calculus.) Since DD^T is invertible, the optimality condition (11) can be written as

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

The maximum fitting error bound in (5) follows from the optimality condition above. For any $\nu \in \mathbf{R}^{n-2}$ with $\nu_t \in [-\lambda, \lambda]$,

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

It follows from (11) that the minimizer x of (4) satisfies

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

We can now derive the formula (6) for $\lambda_{\max}$. Since x^{ba} is affine, $Dx^{\mathrm{ba}} = 0$, so the condition that x^{ba} is optimal is that $\big((DD^T)^{-1}D(y - x^{\mathrm{ba}})\big)_t \in [-\lambda, \lambda]$ for $t = 1, \ldots, n-2$, *i.e.*,

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

5.2 Dual problem

To derive a Lagrange dual of the primal problem of minimizing (4), 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

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

The dual function is

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

The dual problem is

$$\begin{array}{ll} \text{minimize} & g(\nu) = (1/2)\nu^T DD^T\nu - y^T\nu \\ \text{subject to} & -\lambda\mathbf{1} \le \nu \le \lambda\mathbf{1}. \end{array} \tag{12}$$

(Here $a \le b$ means $a_i \le b_i$ for all i.) The dual problem (12) 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 (12), we can recover the ℓ_1 trend estimate via

$$x^{\mathrm{lt}} = y - D^T\nu^{\mathrm{lt}}. \tag{13}$$

6 A primal-dual interior point method

The QP (12) can be solved by standard convex optimization methods, including general interior-point methods [10, 51, 73], and more specialized methods such as path following [53, 26]. 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.*, [10, §11.7] or [73].

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 (12), is $O(n)$. It follows that the over all complexity is $O(n)$, the same as 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

$$r_t(\nu, \mu_1, \mu_2) = 0, \tag{14}$$

where $t > 0$ is a parameter and

$$r_t(\nu,\mu_1,\mu_2) \;=\; \begin{bmatrix} r_{\mathrm{dual}} \\ r_{\mathrm{cent}} \end{bmatrix} = \begin{bmatrix} \nabla g(\nu) + D(\nu - \lambda\mathbf{1})^T\mu_1 - D(\nu+\lambda\mathbf{1})^T\mu_2 \\ -\mu_1(\nu - \lambda\mathbf{1}) + \mu_2(\nu + \lambda\mathbf{1}) - (1/t)\mathbf{1} \end{bmatrix} \tag{15}$$

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 (12), 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 (12). 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 sequenece 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

$$\begin{bmatrix} DD^T & I & -I \\ I & J_1 & 0 \\ -I & 0 & J_2 \end{bmatrix} \begin{bmatrix} \Delta\nu \\ \Delta\mu_1 \\ \Delta\mu_2 \end{bmatrix} = - \begin{bmatrix} DD^T z - Dy + \mu_1 - \mu_2 \\ f_1 + (1/t)\,\mathrm{diag}(\mu_1)^{-1}\mathbf{1} \\ f_2 + (1/t)\,\mathrm{diag}(\mu_2)^{-1}\mathbf{1} \end{bmatrix}, \tag{16}$$

where

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

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

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

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

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

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

A simple Matlab 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. This Matlab implementation scales well to problems with one million data points, which it solves in around a few minutes.

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 handled 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:

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

where the variables are the local trend parameters α_k and β_k. This technique is described (in another context) in, *e.g.*, [10, §6.5].

7.2 Iterative weighted ℓ_1 heuristic

The basic ℓ_1 trend filtering method is equivalent to

$$\begin{array}{ll} \text{minimize} & \|Dx\|_1 \\ \text{subject to} & \|y - x\|_2 \le s, \end{array} \tag{17}$$

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:

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

where $\mathbf{card}(z)$ 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 $\mathbf{card}(Dx)$ with $\|Dx\|_1$, which gives us our basic ℓ_1 trend filter, *i.e.*, the solution to (17). 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 (17). 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

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

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 $\mathbf{card}(Dx)$, 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 [48], where an interpretation of the heuristic for cardinality minimization is given.

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 (4) subject to this monotonicity constraint is an extension of isotonic regression, which has been extensively studied in statistics [3, 58]. (Related work on ℓ_1-regularized isotonic regression, in an engineering context, includes [30, 31].)

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)

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

For example, using $\psi(u) = |u|$, we assign less penalty (compared to $\psi(u) = (1/2)u^2$) to large residuals, but 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 [38]. 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 [40] for more on the use of this penalty function in quantile regression and [41] 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.*, [9, 25, 33, 34] and 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 (4), 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 (4) 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:

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

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.*, [12, 25, 39, 45, 60].)

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

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

where the variables are x (the estimated trend) and s (the estimated seasonal component). The last equality constraint means that the season component sums to zero over the period; without this constraint, the decomposition is not unique [29, §6.2.8]. (For simplicity, we show here the case $n = mp$.) To smooth the seasonal component, we can add a penalty term to the objective, or impose a constraint on the variation of s.

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 is simplified as

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

where the variables are $x \in \mathbf{R}^n$ and $a, b \in \mathbf{R}$.

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 [70].) 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 compressed sensing involving complex quantities and related to total variation in two-dimensional data; see [14].) 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 four 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 trend shifts in slope.

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.

We first consider a penalized least squares problem of the form

$$\text{minimize} \quad (1/2)\textstyle\sum_{i=1}^{n}(y_i - x(t_i))^2 + \lambda \int_{t_1}^{t_n} (\ddot{x}(t))^2\, dt, \qquad (18)$$

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.*, [28, 72]. 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

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

then the objective of the continuous-time problem (18) 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

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

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 [55] for more on this space.) The difference from (18) is that in the integral term the second derivative is penalized using the absolute value function.

A standard result in interpolation theory [55] is that the solution of the interpolation problem

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

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 (19) 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 $\delta(t)$ 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

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

We have seen that the continuous ℓ_1 filtering problem (19) is equivalent to the convex problem

$$\text{minimize} \quad (1/2)\sum_{i=1}^n (y_i - x_i)^2 + \lambda \sum_{i=2}^{n-1} \left| \frac{x_{i+1} - x_i}{t_{i+1} - t_i} - \frac{x_i - x_{i-1}}{t_i - t_{i-1}} \right| \tag{20}$$

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 (19). When t_i are regularly spaced, this problem reduces to the basic ℓ_1 trend filtering problem considered in §3. For the same reason, we can solve (20) (and hence (19)) in $O(n)$ arithmetic operations.

Acknowledgments

The authors thank Johan Lim, Michael Lustig, and Almir Mutapcic for helpful comments and suggestions. This work was funded in part by the Focus Center Research Program Center for Circuit & System Solutions (www.c2s2.org), under contract 2003-CT-888, by AFOSR grant AF F49620-01-1-0365, by NSF grant ECS-0423905, by NSF grant 0529426, by DARPA/MIT grant 5710001848, by AFOSR grant FA9550-06-1-0514, DARPA/Lockheed contract N66001-06-C-2021, and by AFOSR/Vanderbilt grant FA9550-06-1-0312.

References

[1] R. Baillie and S. Chung. Modeling and forecasting from trend stationary long memory models with applications to climatology. *International Journal of Forecasting*, 18(2):215–226, 2002.

[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*. Wiley, New York, 1972.

[4] M. Baxter and R. King. Measuring business cycles: approximate band-pass filters for economic time series. *The Review of Economics and Statistics*, 81(4):575–593, 1999.

[5] D. Bertsekas. *Nonlinear Programming*. Athena Scientific, second edition, 1999.

[6] P. Bloomfeld. Trends in global temperature. *Climate Change*, 21:1–16, 1992.

[7] P. Bloomfeld and D. Nychka. Climate spectra and detecting climate change. *Climate Change*, 21:275–287, 1992.

[8] J. Borwein and A. Lewis. *Convex Analysis and Nonlinear Optimization*. Springer, 2000.

[9] G. Box, G. Jenkins, and G. Reinsel. *Time Series Analysis: Forecasting & Control*. Prentice Hall, third edition, 1994.

[10] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[11] S. Boyd, L. Vandenberghe, A. El Gamal, and S. Yun. Design of robust global power and ground networks. In *Proceedings of ACM/SIGDA International Symposium on Physical Design (ISPD)*, pages 60–65, 2001.

[12] J. Burman. Seasonal adjustment by signal extraction. *Journal of the Royal Statistical Society, Series A*, 143:321–337, 1980.

[13] E. Candès, J. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59(8):1207–1223, 2005.

[14] E. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, 2006.

[15] E. Candès and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005.

[16] S. Chen and D. Donoho. Basis pursuit. In *Proceedings of the Twenty-Eighth Asilomar Conference on Signals, Systems and Computers*, volume 1, pages 41–44, 1994.

[17] S. Chen, D. Donoho, and M. Saunders. Atomic decomposition by basis pursuit. *SIAM Review*, 43(1):129–159, 2001.

[18] L. Christiano and T. Fitzgerald. The band-pass filter. *International Economic Review*, 44(2):435–465, 2003.

[19] P. Craigmile, P. Guttorp, and D. Percival. Trend assessment in a long memory dependence model using the discrete wavelet transform. *Environmetrics*, 15(4):313–335, 2004.

[20] J. Dahl, V. Roychowdhury, and L. Vandenberghe. Maximum likelihood estimation of Gaussian graphical models: Numerical implementation and topology selection. Submitted. Available from `www.ee.ucla.edu/~vandenbe/covsel.html`, 2005.

[21] A. d'Aspremont, L. El Ghaoui, M. Jordan, and G. Lanckriet. A direct formulation for sparse PCA using semidefinite programming, 2005. In L. Saul, Y. Weiss and L. Bottou, editors, *Advances in Neural Information Processing Systems*, 17, pp. 41-48, MIT Press.

[22] D. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

[23] D. Donoho, M. Elad, and V. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Transactions On Information Theory*, 52(1):6–18, 2006.

[24] D. Donoho, I. Johnstone, G. Kerkyacharian, and D. Picard. Wavelet shrinkage: Asymptopia? *J. R. Statist. Soc. B.*, 57(2):301–337, 1995.

[25] J. Durbin and S. Koopman. *Time Series Analysis by State Space Methods*. Oxford University Press, 2001.

[26] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32(2):407–499, 2004.

[27] M. Elad, J. Starck, D. Donoho, and P. Querre. Simultaneous cartoon and texture image inpainting using morphological component analysis (MCA). *Applied and Computational Harmonic Analysis*, 19:340–358, 2005.

[28] R. Eubank. *Nonparametric Regression and Spline Smoothing*. Marcel Dekker, 1999.

[29] J. Fan and Q. Yao. *Nonlinear Time Series: Nonparametric and Parametric Methods*. Springer-Verlag, New York, 2003.

[30] D. Gorinevsky. Monotonic regression filters for trending gradual deterioration faults. In *Proceedings of American Control Conference (ACC)*, pages 5394–5399, 2004.

[31] D. Gorinevsky, S.-J. Kim, S. Boyd, G. Gordon, S. Beard, and F.-K. Chang. Optimal estimation of accumulating damage trend from a series of SHM images, 2007. Accepted to *the 6th International Workshop on Structural Health Monitoring*.

[32] M. Grant, S. Boyd, and Y. Ye. `cvx`*: a Matlab Software for Disciplined Convex Programming*, 2007. Available from `www.stanford.edu/~boyd/cvx/`.

[33] J. Hamilton. *Time Series Analysis*. Princeton University Press, 1994.

[34] A. Harvey. *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press, 1991.

[35] A. Hassibi, J. How, and S. Boyd. Low-authority controller design via convex optimization. In *Proceedings of the IEEE Conference on Decision and Control*, pages 140–145, 1999.

[36] T. Hastie, J. Taylor, R. Tibshirani, and G. Walther. Forward stagewise regression and the monotone lasso. *Electronic Journal of Statistics*, 1:1–29, 2007.

[37] R. Hodrick and E. Prescott. Postwar U.S. business cycles: An empirical investigation. *Journal of Money, Credit, and Banking*, 29(1):1–16, 1997. Discussion paper No. 451. Center for Mathematical Studies in Economics and Management Science, Northwestern University, Evanston, IL, 1981.

[38] P. J. Huber. *Robust Statistics*. John Wiley & Sons, 1981.

[39] P. Kenny and J. Durbin. Local trend estimation and seasonal adjustment of economic and social time series. *Journal of Royal Statistical Society. Series A*, 145:1–41, 1982.

[40] R. Koenker and G. Bassett. Regression quantile. *Econometrica*, 46(1):33–50, 1978.

[41] R. Koenker, P. Ng, and S. Portnoy. Quantile smoothing splines. *Biometrika*, 81(4):673–680, 1994.

[42] K. Koh, S.-J. Kim, and S. Boyd. An interior-point method for ℓ_1-regularized logistic regression, 2006. To appear at *the 22nd National Conference on Artificial Intelligence (AAAI-07)*. Available from `www.stanford.edu/~boyd/l1_logistic_reg.html`.

[43] 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.

[44] C. Leser. A simple method of trend construction. *Journal of the Royal Statistical Society. Series B (Methodological)*, 23(1):91–107, 1961.

[45] C. Leser. Estimation of quasi-linear trend and seasonal variation. *Journal of the American Statistical Association*, 58(304):1033–1043, 1963.

[46] S. Levitt. Understanding why crime fell in the 1990s: Four factors that explain the decline and six that do not. *Journal of Economic Perspectives*, 18(1):163–190, 2004.

[47] W. Link and J. Sauer. Estimating equations estimates of trend. *Bird Populations*, 2:23–32, 1994.

[48] M. Lobo, M. Fazel, and S. Boyd. Portfolio optimization with linear and fixed transaction costs. *Annals of Operations Research*, 152(1):341–365, 2006.

[49] R. Lucas. Two illustrations of the quantity theory of money. *American Economic Review*, 70:1005–14, 1980.

[50] G. Mosheiov and A. Raveh. On trend estimation of time-series: A simpler linear programming approach. *Journal of the Operations Research Society*, 48(1):90–96, 1997.

[51] Y. Nesterov and A. Nemirovsky. *Interior-Point Polynomial Methods in Convex Programming*, volume 13 of *Studies in Applied Mathematics*. SIAM, Philadelphia, PA, 1994.

[52] D. Osborne. Moving average detrending and the analysis of business cycles. *Oxford Bulletin of Economics and Statistics*, 57:547–558, 1995.

[53] M. Osborne, B. Presnell, and B. Turlach. A new approach to variable selection in least squares problems. *IMA Journal of Numerical Analysis*, 20(3):389–403, 2000.

[54] M. Park and T. Hastie. An ℓ_1 regularization-path algorithm for generalized linear models, 2006. To appear in *Journal of the Royal Statistical Society, Series B*.

[55] A. Pinkus. On smoothest interpolants. *SIAM Journal on Mathematical Analysis*, 19(6):1431–41, 1988.

[56] D. Pollock. Trend estimation and de-trending via rational square-wave filters. *Journal of Econometrics*, 99(2):317–334, 2000.

[57] C. Reinsch. Smoothing by spline functions. *Numerische Mathematik*, 10:177–183, 1976.

[58] T. Robertson, F. Wright, R., and Dykstra. *Order Restricted Statistical Inference*. Wiley, New York, 1988.

[59] S. Rosset and J. Zhu. Piecewise linear regularized solution paths. *Annals of Statistics*, 35(3), 2007.

[60] E. Schlicht. A seasonal adjustment principle and a seasonal adjustment method derived from this principle. *Journal of the American Statistical Association*, 76(374):374–378, 1981.

[61] K. Singleton. Econometric issues in the analysis of equilibrium business cycle models. *Journal of Monetary Economics*, 21(2-3):361–386, 1988.

[62] J. Starck, M. Elad, and D. Donoho. Image decomposition via the combination of sparse representations and a variational approach. *IEEE Transactions on Image Processing*, 14(10):1570–1582, 2005.

[63] J. Sturm. *Using SEDUMI 1.02, a Matlab Toolbox for Optimization Over Symmetric Cones*, 2001. Available from `fewcal.kub.nl/sturm/software/sedumi.html`.

[64] R. Talluri and G. van Ryzin. *Theory and Practice of Revenue Management*. Springer, 2005.

[65] R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288, 1996.

[66] R. Tibshirani, M. Saunders, S. Rosset, and J. Zhu. Sparsity and smoothness via the fused Lasso. *Journal of the Royal Statistical Society Series B*, 67(1):91–108, 2005.

[67] J. Tropp. Just relax: Convex programming methods for identifying sparse signals in noise. *IEEE Transactions on Information Theory*, 53(3):1030–1051, 2006.

[68] R. Tsay. *Analysis of Financial Time Series*. Wiley-Interscience, second edition, 2005.

[69] M. Wainwright, P. Ravikumar, and J. Lafferty. High-dimensional graphical model selection using ℓ_1-regularized logistic regression., 2007. To appear in *Advances in Neural Information Processing Systems (NIPS) 19*. Available from `http://www.eecs.berkeley.edu/~wainwrig/Pubs/publist.html#High-dimensional`.

[70] J. Wang. A Bayesian time series model of multiple structural changes in level, trend, and variance. *Journal of Busines and Economic Statistics*, 18(3):374–386, 2000.

[71] Y. Wen and B. Zeng. A simple nonlinear filter for economic time series analysis. *Economics Letters*, 64:151160, 1999.

[72] G. Whaba. *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics, 1990.

[73] S. Wright. *Primal-dual interior-point methods*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1997.

[74] S. Zhao and G. Wei. Jump process for the trend estimation of time series. *Computational Statistics & Data Analysis*, 42:219–241, 2003.

[75] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67(2):301–320, 2005.