Wavelet-Based Signal Extraction and Denoising

- overview of key ideas behind wavelet-based approach
- description of four basic models for signal estimation
- discussion of why wavelets can help estimate certain signals
- simple thresholding & shrinkage schemes for signal estimation
- wavelet-based thresholding and shrinkage
- case studies:
  - denoising ECG time series
  - spectral density function estimation (if time permits)
    - wavelet-based approach using periodogram
    - wavelet-based approach using multitaper estimators
- brief comments on 'second generation' denoising

Wavelet-Based Signal Estimation: I

- DWT analysis of $\mathbf{X}$ yields $\mathbf{W} = \mathcal{W}\mathbf{X}$
- DWT synthesis $\mathbf{X} = \mathcal{W}^T\mathbf{W}$ yields multiresolution analysis by splitting $\mathcal{W}^T\mathbf{W}$ into pieces associated with different scales
- DWT synthesis can also estimate 'signal' hidden in $\mathbf{X}$ if we can modify $\mathbf{W}$ to get rid of noise in the wavelet domain
- if $\mathbf{W}'$ is a 'noise reduced' version of $\mathbf{W}$, can form signal estimate via $\mathcal{W}^T\mathbf{W}'$

Wavelet-Based Signal Estimation: II

- key ideas behind simple wavelet-based signal estimation
  - certain signals can be efficiently described by the DWT using
    - all of the scaling coefficients
    - a small number of 'large' wavelet coefficients
  - noise is manifested in a large number of 'small' wavelet coefficients
  - can either 'threshold' or 'shrink' wavelet coefficients to eliminate noise in the wavelet domain
- key ideas led to wavelet thresholding and shrinkage proposed by Donoho, Johnstone and coworkers in 1990s

Models for Signal Estimation: I

- will consider two types of signals:
  1. $\mathbf{D}$, an N dimensional deterministic signal
  2. $\mathbf{C}$, an N dimensional stochastic signal; i.e., a vector of random variables (RVs) with covariance matrix $\Sigma_{\mathbf{C}}$
- will consider two types of noise:
  1. $\boldsymbol{\epsilon}$, an N dimensional vector of independent and identically distributed (IID) RVs with mean 0 and covariance matrix $\Sigma_{\boldsymbol{\epsilon}} = \sigma_{\epsilon}^2 I_N$
  2. $\boldsymbol{\eta}$, an N dimensional vector of non-IID RVs with mean 0 and covariance matrix $\Sigma_{\boldsymbol{\eta}}$
     * one form: RVs independent, but have different variances
     * another form of non-IID: RVs are correlated

Models for Signal Estimation: II

- leads to four basic 'signal + noise' models for $\mathbf{X}$
  1. $\mathbf{X} = \mathbf{D} + \boldsymbol{\epsilon}$
  2. $\mathbf{X} = \mathbf{D} + \boldsymbol{\eta}$
  3. $\mathbf{X} = \mathbf{C} + \boldsymbol{\epsilon}$
  4. $\mathbf{X} = \mathbf{C} + \boldsymbol{\eta}$
- in the latter two cases, the stochastic signal $\mathbf{C}$ is assumed to be independent of the associated noise

Signal Representation via Wavelets: I

- consider deterministic signals $\mathbf{D}$ first
- signal estimation problem is simplified if we can assume that the important part of $\mathbf{D}$ is in its large values
- assumption is not usually viable in the original (i.e., time domain) representation $\mathbf{D}$, but might be true in another domain
- an orthonormal transform $\mathcal{O}$ might be useful because
  - $\mathbf{O} = \mathcal{O}\mathbf{D}$ is equivalent to $\mathbf{D}$ (since $\mathbf{D} = \mathcal{O}^T\mathbf{O}$)
  - we might be able to find $\mathcal{O}$ such that the signal is isolated in $M \ll N$ large transform coefficients
- Q: how can we judge whether a particular $\mathcal{O}$ might be useful for representing $\mathbf{D}$?

Signal Representation via Wavelets: II

- let O_j be the jth transform coefficient in $\mathbf{O} = \mathcal{O}\mathbf{D}$
- let $O_{(0)}, O_{(1)}, \ldots, O_{(N-1)}$ be the O_j's reordered by magnitude:

$$|O_{(0)}| \geq |O_{(1)}| \geq \cdots \geq |O_{(N-1)}|$$

- example: if $\mathbf{O} = [-3, 1, 4, -7, 2, -1]^T$, then $O_{(0)} = O_3 = -7$, $O_{(1)} = O_2 = 4$, $O_{(2)} = O_0 = -3$ etc.
- define a normalized partial energy sequence (NPES):

$$C_{M-1} \equiv \frac{\sum_{j=0}^{M-1} |O_{(j)}|^2}{\sum_{j=0}^{N-1} |O_{(j)}|^2} = \frac{\text{energy in largest } M \text{ terms}}{\text{total energy in signal}}$$

- let $\mathcal{I}_M$ be $N \times N$ diagonal matrix whose jth diagonal term is 1 if $|O_j|$ is one of the M largest magnitudes and is 0 otherwise

Signal Representation via Wavelets: III

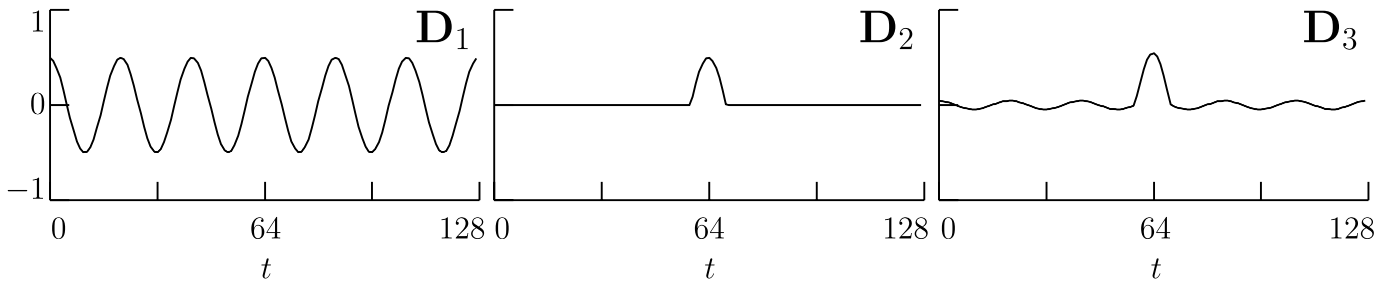
- form $\widehat{\mathbf{D}}_M \equiv \mathcal{O}^T \mathcal{I}_M \mathbf{O}$, which is an approximation to $\mathbf{D}$
- when $\mathbf{O} = [-3, 1, 4, -7, 2, -1]^T$ and $M = 3$, we have

$$\mathcal{I}_3 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad \text{and thus} \quad \widehat{\mathbf{D}}_M = \mathcal{O}^T \begin{bmatrix} -3 \\ 0 \\ 4 \\ -7 \\ 0 \\ 0 \end{bmatrix}$$

- Exer. [395] shows that

$$C_{M-1} = 1 - \frac{\|\mathbf{D} - \widehat{\mathbf{D}}_M\|^2}{\|\mathbf{D}\|^2} = 1 - \text{relative approximation error}$$

Signal Representation via Wavelets: IV



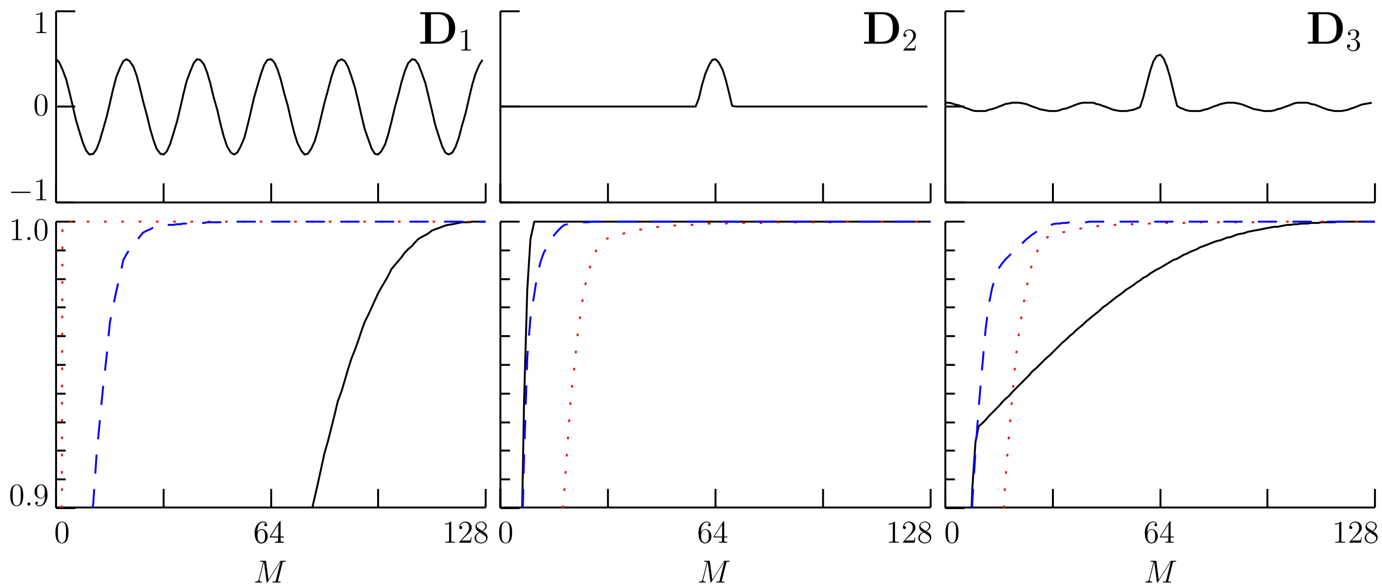
- consider three signals plotted above
- $\mathbf{D}_1$ is a sinusoid, which can be represented succinctly by the discrete Fourier transform (DFT)
- $\mathbf{D}_2$ is a bump (only a few nonzero values in the time domain)
- $\mathbf{D}_3$ is a linear combination of $\mathbf{D}_1$ and $\mathbf{D}_2$

Signal Representation via Wavelets: V

- consider three different orthogonal transforms
  - identity transform I (time)
  - the orthogonal DFT $\mathcal{F}$ (frequency), where $\mathcal{F}$ has (k, t)th element $\exp(-i2\pi tk/N)/\sqrt{N}$ for $0 \leq k, t \leq N-1$
  - the LA(8) DWT $\mathcal{W}$ (wavelet)
- # of terms M needed to achieve relative error $< 1\%$:

	$\mathbf{D}_1$	$\mathbf{D}_2$	$\mathbf{D}_3$
DFT	2	29	28
identity	105	9	75
LA(8) wavelet	22	14	21

Signal Representation via Wavelets: VI

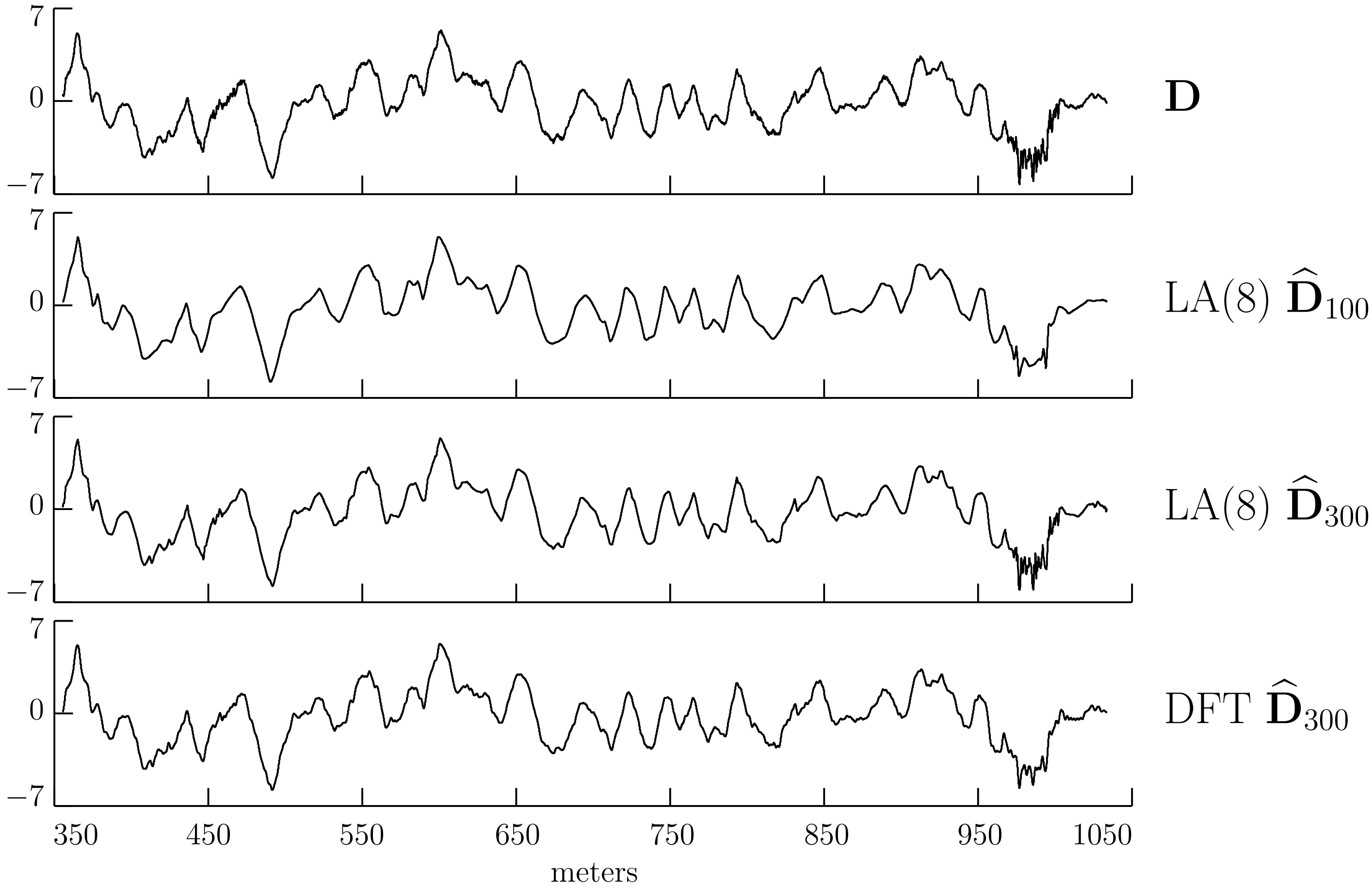


- use NPESs to see how well these three signals are represented in the time, frequency (DFT) and wavelet (LA(8)) domains
- time (solid curves), frequency (dotted) and wavelet (dashed)

Signal Representation via Wavelets: VII

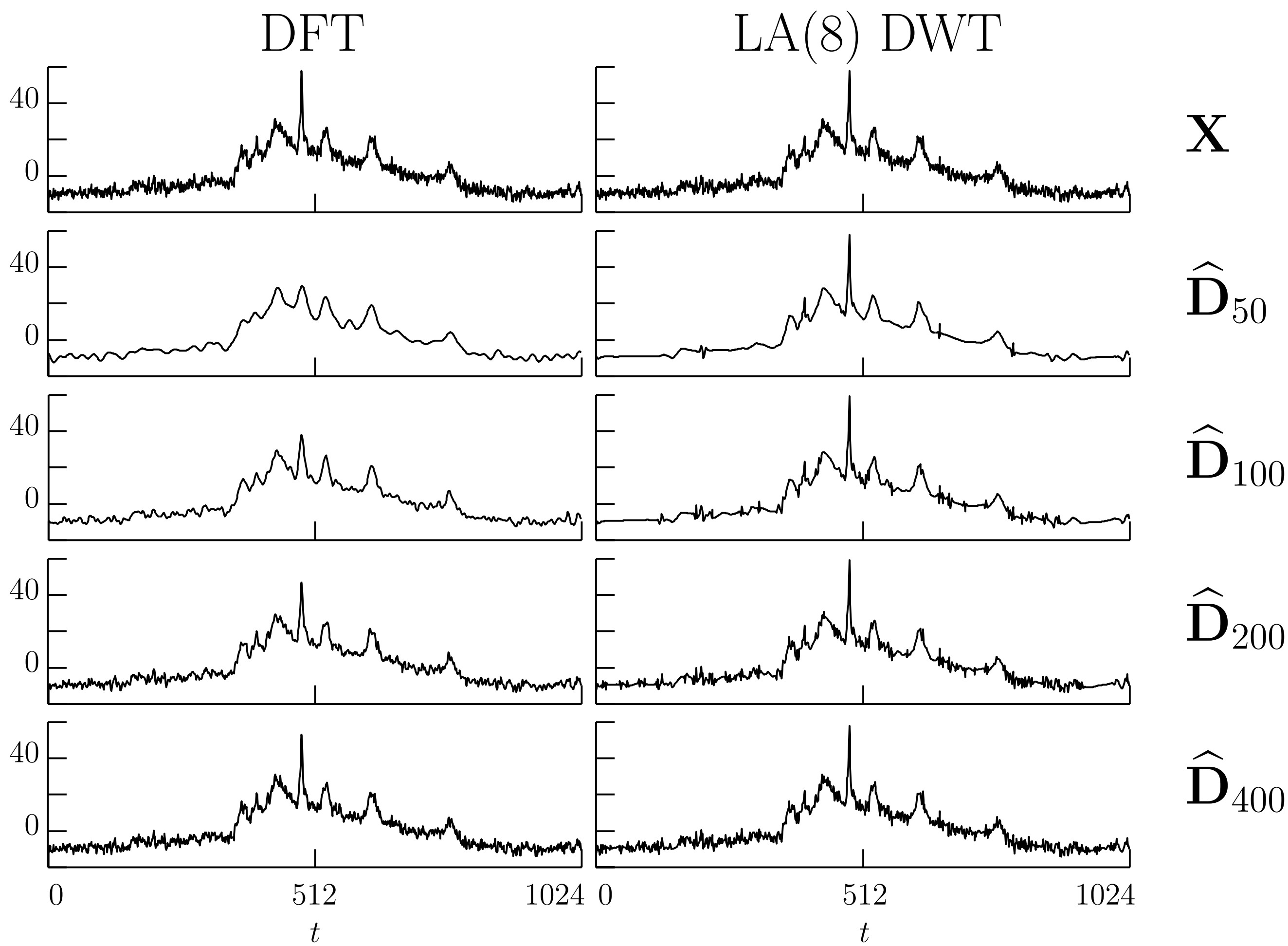
- let us consider the vertical ocean shear time series as a 'signal'
- will look at plots of
  - the signal $\mathbf{D}$ itself
  - its approximation $\widehat{\mathbf{D}}_{100}$ from 100 LA(8) DWT coefficients
  - $\widehat{\mathbf{D}}_{300}$ from 300 LA(8) DWT coefficients, giving $C_{299} \doteq 0.9983$
  - $\widehat{\mathbf{D}}_{300}$ from 300 DFT coefficients, giving $C_{299} \doteq 0.9973$
- note that 300 coefficients is less than 5% of $N = 6784$!

Signal Representation via Wavelets: VIII



- need 123 additional DFT coefficients to match C_{299} for DWT

Signal Representation via Wavelets: IX



- 2nd example: DFT $\widehat{\mathbf{D}}_M$ (left-hand column) & $J_0 = 6$ LA(8) DWT $\widehat{\mathbf{D}}_M$ (right) for NMR series $\mathbf{X}$ (A. Maudsley, UCSF)

Signal Estimation via Thresholding: I

- assume model of deterministic signal plus IID noise: $\mathbf{X} = \mathbf{D} + \boldsymbol{\epsilon}$
- let $\mathcal{O}$ be an $N \times N$ orthonormal matrix
- form $\mathbf{O} = \mathcal{O}\mathbf{X} = \mathcal{O}\mathbf{D} + \mathcal{O}\boldsymbol{\epsilon} \equiv \mathbf{d} + \mathbf{e}$
- component-wise, have $O_l = d_l + e_l$
- define signal to noise ratio (SNR):

$$\frac{\|\mathbf{D}\|^2}{E\{\|\boldsymbol{\epsilon}\|^2\}} = \frac{\|\mathbf{d}\|^2}{E\{\|\mathbf{e}\|^2\}} = \frac{\sum_{l=0}^{N-1} d_l^2}{\sum_{l=0}^{N-1} E\{e_l^2\}}$$

- assume that SNR is large
- assume that $\mathbf{d}$ has just a few large coefficients; i.e., large signal coefficients dominate $\mathbf{O}$

Signal Estimation via Thresholding: II

- recall simple estimator $\widehat{\mathbf{D}}_M \equiv \mathcal{O}^T \mathcal{I}_M \mathbf{O}$ and previous example:

$$\widehat{\mathbf{D}}_M = \mathcal{O}^T \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} O_0 \\ O_1 \\ O_2 \\ O_3 \\ O_4 \\ O_5 \end{bmatrix} = \mathcal{O}^T \begin{bmatrix} O_0 \\ 0 \\ O_2 \\ O_3 \\ 0 \\ 0 \end{bmatrix}$$

- let $\mathcal{J}_m$ be a set of m indices corresponding to places where jth diagonal element of $\mathcal{I}_m$ is 1
- in example above, we have $\mathcal{J}_3 = \{0, 2, 3\}$
- strategy in forming $\widehat{\mathbf{D}}_M$ is to keep a coefficient O_j if $j \in \mathcal{J}_m$ but to replace it with 0 if $j \notin \mathcal{J}_m$ ('kill' or 'keep' strategy)

Signal Estimation via Thresholding: III

- can pose a simple optimization problem whose solution
  1. is a 'kill or keep' strategy (and hence justifies this strategy)
  2. dictates that we use coefficients with the largest magnitudes
  3. tells us what M should be (once we set a certain parameter)
- optimization problem: find $\widehat{\mathbf{D}}_M$ such that

$$\gamma_m \equiv \|\mathbf{X} - \widehat{\mathbf{D}}_m\|^2 + m\delta^2$$

  is minimized over all possible $\mathcal{I}_m$, $m = 0, \ldots, N$
- in the above δ^2 is a fixed parameter (set *a priori*)

Signal Estimation via Thresholding: IV

- $\|\mathbf{X} - \widehat{\mathbf{D}}_m\|^2$ is a measure of 'fidelity'
  - rationale for this term: under our assumption of a high SNR, $\widehat{\mathbf{D}}_m$ shouldn't stray too far from $\mathbf{X}$
  - fidelity increases (the measure decreases) as m increases
  - in minimizing γ_m, consideration of this term alone suggests that m should be large
- $m\delta^2$ is a penalty for too many terms
  - rationale: heuristic says $\mathbf{d}$ has only a few large coefficients
  - penalty increases as m increases
  - in minimizing γ_m, consideration of this term alone suggests that m should be small
- optimization problem: balance off fidelity & parsimony

Signal Estimation via Thresholding: V

- claim: $\gamma_m = \|\mathbf{X} - \widehat{\mathbf{D}}_m\|^2 + m\delta^2$ is minimized when m is set to the number of coefficients O_j such that $O_j^2 > \delta^2$
- proof of claim: since $\mathbf{X} = \mathcal{O}^T\mathbf{O}$ & $\widehat{\mathbf{D}}_m \equiv \mathcal{O}^T\mathcal{I}_m\mathbf{O}$, have

$$\begin{aligned}
\gamma_m = \|\mathbf{X} - \widehat{\mathbf{D}}_m\|^2 + m\delta^2 &= \|\mathcal{O}^T\mathbf{O} - \mathcal{O}^T\mathcal{I}_m\mathbf{O}\|^2 + m\delta^2 \\
&= \|\mathcal{O}^T(I_N - \mathcal{I}_m)\mathbf{O}\|^2 + m\delta^2 \\
&= \|(I_N - \mathcal{I}_m)\mathbf{O}\|^2 + m\delta^2 \\
&= \sum_{j \notin \mathcal{J}_m} O_j^2 + \sum_{j \in \mathcal{J}_m} \delta^2
\end{aligned}$$

- for any given j, if $j \notin \mathcal{J}_m$, we contribute O_j^2 to first sum; on the other hand, if $j \in \mathcal{J}_m$, we contribute δ^2 to second sum
- to minimize γ_m, we need to put j in $\mathcal{J}_m$ if $O_j^2 > \delta^2$, thus establishing the claim

Thresholding Functions: I

- more generally, thresholding schemes involve

  1. computing $\mathbf{O} \equiv \mathcal{O}\mathbf{X}$
  2. defining $\mathbf{O}^{(t)}$ as vector with lth element

  $$O_l^{(t)} = \begin{cases} 0, & \text{if } |O_l| \leq \delta; \\ \text{some nonzero value}, & \text{otherwise}, \end{cases}$$

  where nonzero values are yet to be defined

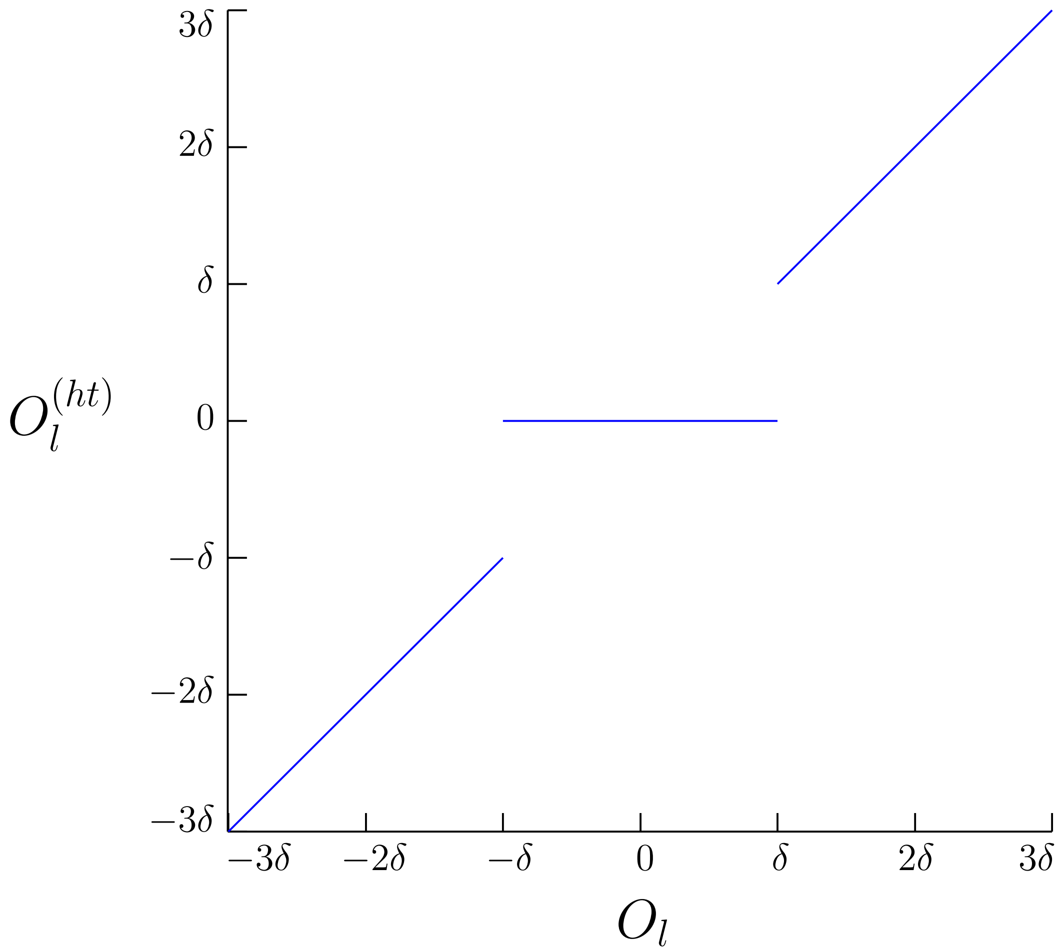
  3. estimating $\mathbf{D}$ via $\widehat{\mathbf{D}}^{(t)} \equiv \mathcal{O}^T\mathbf{O}^{(t)}$

- simplest scheme is 'hard thresholding' ('kill/keep' strategy):

$$O_l^{(ht)} = \begin{cases} 0, & \text{if } |O_l| \leq \delta; \\ O_l, & \text{otherwise.} \end{cases}$$

Thresholding Functions: II

- plot shows mapping from O_l to $O_l^{(ht)}$



Thresholding Functions: III

- alternative scheme is 'soft thresholding:'

$$O_l^{(st)} = \text{sign}\,\{O_l\}\,(|O_l| - \delta)_+\,,$$

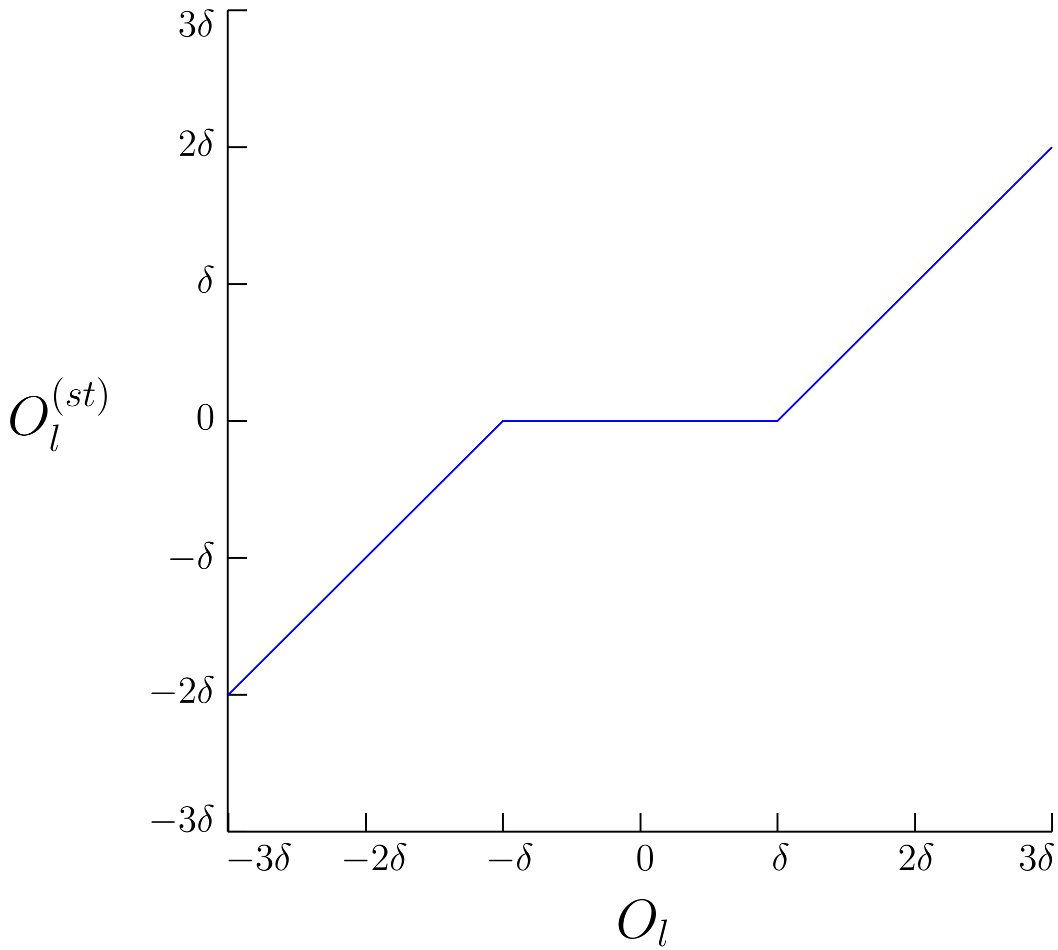
  where

$$\text{sign}\,\{O_l\} \equiv \begin{cases} +1, & \text{if } O_l > 0; \\ 0, & \text{if } O_l = 0; \\ -1, & \text{if } O_l < 0. \end{cases} \quad \text{and} \quad (x)_+ \equiv \begin{cases} x, & \text{if } x \geq 0; \\ 0, & \text{if } x < 0. \end{cases}$$

- one rationale for soft thresholding is that it fits into Stein's class of estimators (will discuss this later)

Thresholding Functions: IV

- here is the mapping from O_l to $O_l^{(st)}$



Thresholding Functions: V

- third scheme is 'mid thresholding:'

$$O_l^{(mt)} = \operatorname{sign}\{O_l\}\,(|O_l| - \delta)_{++}\,,$$

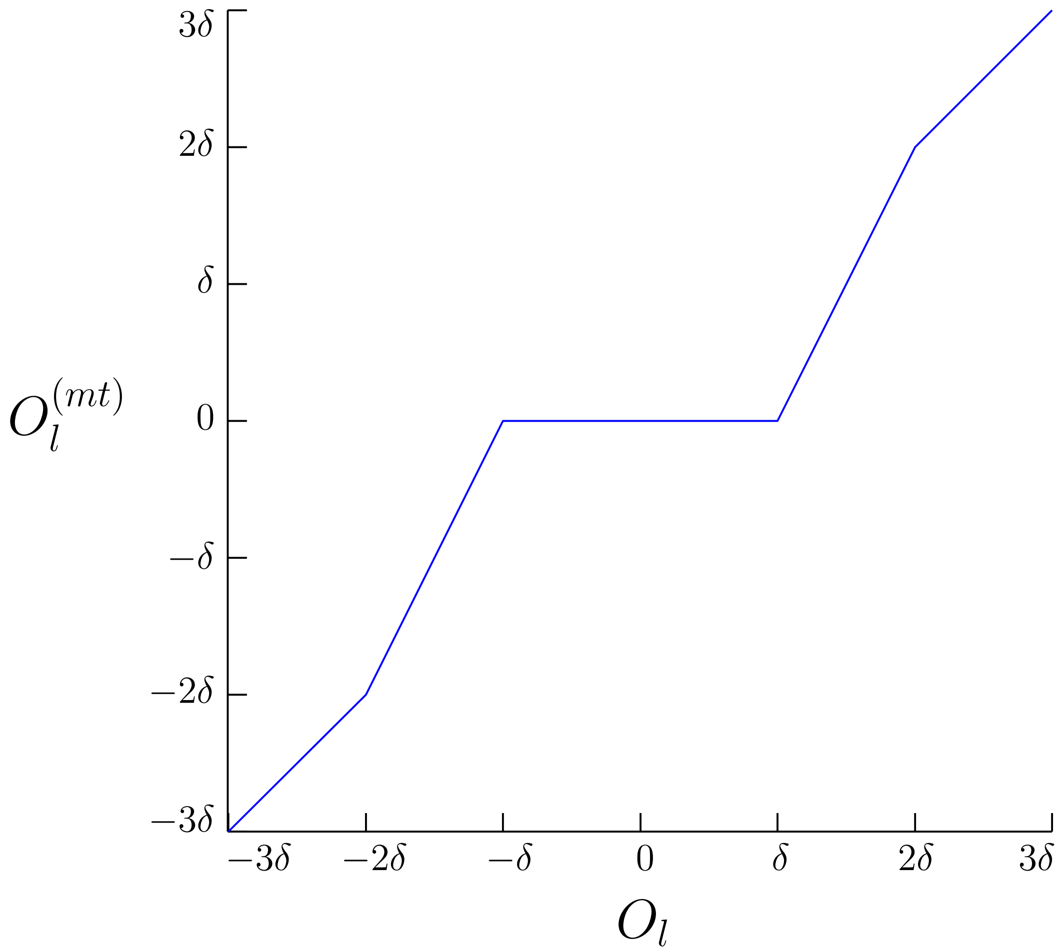
  where

$$(|O_l| - \delta)_{++} \equiv \begin{cases} 2(|O_l| - \delta)_+, & \text{if } |O_l| < 2\delta; \\ |O_l|, & \text{otherwise} \end{cases}$$

- provides compromise between hard and soft thresholding

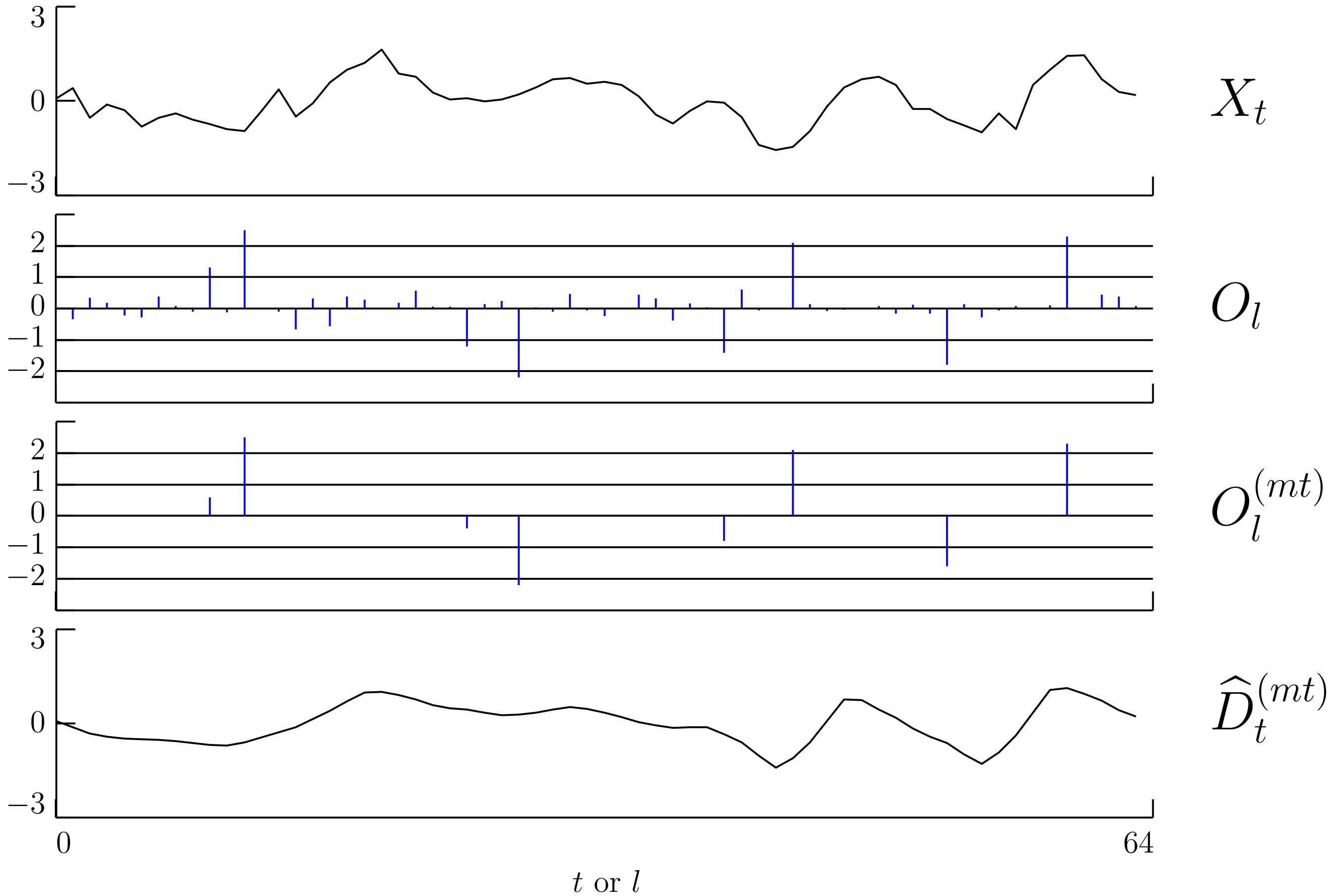
Thresholding Functions: VI

- here is the mapping from O_l to $O_l^{(mt)}$



Thresholding Functions: VII

- example of mid thresholding with $\delta = 1$



Universal Threshold: I

- Q: how do we go about setting δ?
- specialize to IID Gaussian noise $\boldsymbol{\epsilon}$ with covariance $\sigma_\epsilon^2 I_N$
- Exer. [263]: $\mathbf{e} \equiv \mathcal{O}\boldsymbol{\epsilon}$ is also IID Gaussian with covariance $\sigma_\epsilon^2 I_N$
- Donoho & Johnstone (1995) proposed $\delta^{(u)} \equiv \sqrt{[2\sigma_\epsilon^2 \log(N)]}$ ('log' here is 'log base e')
- rationale for $\delta^{(u)}$: because of Gaussianity, can argue that

$$\mathbf{P}\Big[\max_l\{|e_l|\} > \delta^{(u)}\Big] \leq \frac{1}{\sqrt{[4\pi \log{(N)}]}} \to 0 \ \text{ as } \ N \to \infty$$

  and hence $\mathbf{P}\left[\max_l\{|e_l\}| \leq \delta^{(u)}\right] \to 1$ as $N \to \infty$, so no noise will exceed threshold in the limit

Universal Threshold: II

- suppose $\mathbf{D}$ is a vector of zeros so that $O_l = e_l$
- implies that $\mathbf{O}^{(ht)} = 0$ with high probability as $N \to \infty$
- hence will estimate correct $\mathbf{D}$ with high probability
- critique of $\delta^{(u)}$:
  - consider lots of IID Gaussian series, $N = 128$: only 13% will have any values exceeding $\delta^{(u)}$
  - $\delta^{(u)}$ is slanted toward eliminating vast majority of noise, but, if we use, e.g., hard thresholding, any nonzero signal transform coefficient of a fixed magnitude will eventually get set to 0 as $N \to \infty$
- nonetheless: $\delta^{(u)}$ works remarkably well

Minimum Unbiased Risk: I

- second approach for setting δ is data-adaptive, but only works for selected thresholding functions
- assume model of deterministic signal plus non-IID noise: $\mathbf{X} = \mathbf{D} + \boldsymbol{\eta}$ so that $\mathbf{O} \equiv \mathcal{O}\mathbf{X} = \mathcal{O}\mathbf{D} + \mathcal{O}\boldsymbol{\eta} \equiv \mathbf{d} + \mathbf{n}$
- component-wise, have $O_l = d_l + n_l$
- further assume that n_l is an $\mathcal{N}(0, \sigma^2_{n_l})$ RV, where $\sigma^2_{n_l}$ is assumed to be known, but we allow the possibility that n_l's are correlated
- let $O_l^{(\delta)}$ be estimator of d_l based on a (yet to be determined) threshold δ
- put $O_l^{(\delta)}$'s into vector $\mathbf{O}^{(\delta)}$

Minimum Unbiased Risk: II

- define $\widehat{\mathbf{D}}^{(\delta)} \equiv \mathcal{O}^T \mathbf{O}^{(\delta)}$ and associated 'risk'

$$\begin{aligned} R(\widehat{\mathbf{D}}^{(\delta)}, \mathbf{D}) \equiv E\{\|\widehat{\mathbf{D}}^{(\delta)} - \mathbf{D}\|^2\} &= E\{\|\mathcal{O}(\widehat{\mathbf{D}}^{(\delta)} - \mathbf{D})\|^2)\} \\ &= E\{\|\mathbf{O}^{(\delta)} - \mathbf{d}\|^2)\} \\ &= E\left\{ \sum_{l=0}^{N-1} (O_l^{(\delta)} - d_l)^2 \right\} \end{aligned}$$

- can minimize risk by making $E\{(O_l^{(\delta)} - d_l)^2\}$ as small as possible for each l
- Stein (1981) considered estimators restricted to be of the form

$$O_l^{(\delta)} = O_l + A^{(\delta)}(O_l),$$

where $A^{(\delta)}(\cdot)$ must be 'weakly differentiable' (think of it as defining a derivative for a continuous function that is only piecewise differentiable in the usual sense; e.g., soft thresholding)

Minimum Unbiased Risk: III

- using $O_l^{(\delta)} = O_l + A^{(\delta)}(O_l)$ with $O_l = d_l + n_l$ yields

$$O_l^{(\delta)} - d_l = n_l + A^{(\delta)}(O_l)$$

and hence

$$E\{(O_l^{(\delta)} - d_l)^2\} = \sigma_{n_l}^2 + 2E\{n_l A^{(\delta)}(O_l)\} + E\{[A^{(\delta)}(O_l)]^2\}$$

- because of Gaussianity, can reduce middle term:

$$E\{n_l A^{(\delta)}(O_l)\} = \sigma_{n_l}^2 E\left\{ \frac{d}{dx} A^{(\delta)}(x) \bigg|_{x=O_l} \right\}$$

- can now write $E\{(O_l^{(\delta)} - d_l)^2\} = E\{\mathcal{R}(\sigma_{n_l}, O_l, \delta)\}$, where

$$\mathcal{R}(\sigma_{n_l}, x, \delta) \equiv \sigma_{n_l}^2 + 2\sigma_{n_l}^2 \frac{d}{dx} A^{(\delta)}(x) + [A^{(\delta)}(x)]^2$$

Minimum Unbiased Risk: IV

- risk in using $\mathbf{D}^{(\delta)}$ given by

$$R(\widehat{\mathbf{D}}^{(\delta)}, \mathbf{D}) = E\left\{\sum_{l=0}^{N-1} (O_l^{(\delta)} - d_l)^2\right\} = E\left\{\sum_{l=0}^{N-1} \mathcal{R}(\sigma_{n_l}, O_l, \delta)\right\}$$

- practical scheme: given realizations o_l of O_l, find δ minimizing

$$\sum_{l=0}^{N-1} \mathcal{R}(\sigma_{n_l}, o_l, \delta)$$

- for a given δ, above is Stein's unbiased risk estimator (SURE)

Minimum Unbiased Risk: V

- example: if we set

$$A^{(\delta)}(O_l) = \begin{cases} -O_l, & \text{if } |O_l| < \delta; \\ -\delta \operatorname{sign}\{O_l\}, & \text{if } |O_l| \geq \delta, \end{cases}$$

we obtain $O_l^{(\delta)} = O_l + A^{(\delta)}(O_l) = O_l^{(st)}$, i.e., soft thresholding

- for this case, can argue that

$$\mathcal{R}(\sigma_{n_l}, O_l, \delta) = O_l^2 - \sigma_{n_l}^2 + (2\sigma_{n_l}^2 - O_l^2 + \delta^2)1_{[\delta^2,\infty)}(O_l^2),$$

where

$$1_{[\delta^2,\infty)}(x) \equiv \begin{cases} 1, & \text{if } \delta^2 \leq x < \infty; \\ 0, & \text{otherwise.} \end{cases}$$

- only the last term depends on δ, and, as a function of δ, SURE is minimized when last term is minimized

Minimum Unbiased Risk: VI

- data-adaptive scheme is to replace O_l with its realization, say o_l, and to set δ equal to the value, say $\delta^{(S)}$, minimizing
$$\sum_{l=0}^{N-1} (2\sigma_{n_l}^2 - o_l^2 + \delta^2) 1_{[\delta^2,\infty)}(o_l^2),$$
- must have $\delta^{(S)} = |o_l|$ for some l, so minimization is easy
- if n_l have a common variance, i.e., $\sigma_{n_l}^2 = \sigma_0^2$ for all l, need to find minimizer of the following function of δ:
$$\sum_{l=0}^{N-1} (2\sigma_0^2 - o_l^2 + \delta^2) 1_{[\delta^2,\infty)}(o_l^2),$$
(in practice, σ_0^2 is usually unknown, so later on we will consider how to estimate this also)

Signal Estimation via Shrinkage

- so far, we have only considered signal estimation via thresholding rules, which will map some O_l to zeros
- will now consider shrinkage rules, which differ from thresholding only in that nonzero coefficients are mapped to nonzero values rather than exactly zero (but values can be *very* close to zero!)
- there are three approaches that lead us to shrinkage rules
  1. linear mean square estimation
  2. conditional mean and median
  3. Bayesian approach
- will only consider 1 and 2, but one form of Bayesian approach turns out to be identical to 2

Linear Mean Square Estimation: I

- assume model of stochastic signal plus non-IID noise: $\mathbf{X} = \mathbf{C} + \boldsymbol{\eta}$ so that $\mathbf{O} = \mathcal{O}\mathbf{X} = \mathcal{O}\mathbf{C} + \mathcal{O}\boldsymbol{\eta} \equiv \mathbf{R} + \mathbf{n}$
- component-wise, have $O_l = R_l + n_l$
- assume $\mathbf{C}$ and $\boldsymbol{\eta}$ are multivariate Gaussian with covariance matrices $\Sigma_{\mathbf{C}}$ and $\Sigma_{\boldsymbol{\eta}}$
- implies $\mathbf{R}$ and $\mathbf{n}$ are also multivariate Gaussian, but now with covariance matrices $\mathcal{O}\Sigma_{\mathbf{C}}\mathcal{O}^T$ and $\mathcal{O}\Sigma_{\boldsymbol{\eta}}\mathcal{O}^T$
- assume that $E\{R_l\} = 0$ for any component of interest and that R_l & n_l are uncorrelated
- suppose we estimate R_l via a simple scaling of O_l:

$$\widehat{R}_l \equiv a_l O_l, \quad \text{where } a_l \text{ is a constant to be determined}$$

Linear Mean Square Estimation: II

- let us select a_l by making $E\{(R_l - \widehat{R}_l)^2\}$ as small as possible, which, following from Exer. [407], occurs when we set
$$a_l = \frac{E\{R_l O_l\}}{E\{O_l^2\}}$$
- because R_l and n_l are uncorrelated with 0 means and because $O_l = R_l + n_l$, we have
$$E\{R_l O_l\} = E\{R_l^2\} \text{ and } E\{O_l^2\} = E\{R_l^2\} + E\{n_l^2\},$$
yielding
$$\widehat{R}_l = \frac{E\{R_l^2\}}{E\{R_l^2\} + E\{n_l^2\}} O_l = \frac{\sigma_{R_l}^2}{\sigma_{R_l}^2 + \sigma_{n_l}^2} O_l$$
- note: 'optimum' a_l shrinks O_l toward zero, with shrinkage increasing as the noise variance increases

Background on Conditional PDFs: I

- let X and Y be RVs with probability density functions (PDFs) $f_X(\cdot)$ and $f_Y(\cdot)$
- let $f_{X,Y}(x, y)$ be their joint PDF at the point (x, y)
- $f_X(\cdot)$ and $f_Y(\cdot)$ are called marginal PDFs and can be obtained from the joint PDF via integration:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dy$$

- the conditional PDF of Y given $X = x$ is defined as

$$f_{Y|X=x}(y) = \frac{f_{X,Y}(x, y)}{f_X(x)}$$

(read '$|$' as 'given' or 'conditional on')

Background on Conditional PDFs: II

- by definition RVs X and Y are said to be independent if

$$f_{X,Y}(x,y) = f_X(x)f_Y(y),$$

  in which case

$$f_{Y|X=x}(y) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{f_X(x)f_Y(y)}{f_X(x)} = f_Y(y)$$

- thus X and Y are independent if knowing X doesn't allow us to alter our probabilistic description of Y
- $f_{Y|X=x}(\cdot)$ is a PDF, so its mean value is

$$E\{Y|X=x\} = \int_{-\infty}^{\infty} y f_{Y|X=x}(y)\,dy;$$

  the above is called the conditional mean of Y, given X

Background on Conditional PDFs: III

- suppose RVs X and Y are related, but we can only observe X
- suppose we want to approximate the unobservable Y based on some function of the observable X
- example: we observe part of a time series containing a signal buried in noise, and we want to approximate the unobservable signal component based upon a function of what we observed
- suppose we want our approximation to be the function of X, say $U_2(X)$, such that the mean square difference between Y and $U_2(X)$ is as small as possible; i.e., we want

$$E\{(Y - U_2(X))^2\}$$

to be as small as possible

Background on Conditional PDFs: IV

- solution is to use $U_2(X) = E\{Y|X\}$; i.e., the conditional mean of Y given X is our best guess at Y in the sense of minimizing the mean square error (related to fact that $E\{(Y - a)^2\}$ is smallest when $a = E\{Y\}$)
- on the other hand, suppose we want the function $U_1(X)$ such that the mean absolute error $E\{|Y - U_1(X)|\}$ is as small as possible
- the solution now is to let $U_1(X)$ be the conditional median; i.e., we must solve

$$\int_{-\infty}^{U_1(x)} f_{Y|X=x}(y)\, dy = 0.5$$

to figure out what $U_1(x)$ should be when $X = x$

Conditional Mean and Median Approach: I

- assume model of stochastic signal plus non-IID noise:
  $\mathbf{X} = \mathbf{C} + \boldsymbol{\eta}$ so that $\mathbf{O} = \mathcal{O}\mathbf{X} = \mathcal{O}\mathbf{C} + \mathcal{O}\boldsymbol{\eta} \equiv \mathbf{R} + \mathbf{n}$
- component-wise, have $O_l = R_l + n_l$
- because $\mathbf{C}$ and $\boldsymbol{\eta}$ are independent, $\mathbf{R}$ and $\mathbf{n}$ must be also
- suppose we approximate R_l via $\widehat{R}_l \equiv U_2(O_l)$, where $U_2(O_l)$ is selected to minimize $E\{(R_l - U_2(O_l))^2\}$
- solution is to set $U_2(O_l)$ equal to $E\{R_l|O_l\}$, so let's work out what form this conditional mean takes
- to get $E\{R_l|O_l\}$, need the PDF of R_l given O_l, which is

$$f_{R_l|O_l=o_l}(r_l) = \frac{f_{R_l,O_l}(r_l, o_l)}{f_{O_l}(o_l)}$$

Conditional Mean and Median Approach: II

- Exer. [262a]: the joint PDF of R_l and O_l is related to the joint PDF $f_{R_l,n_l}(\cdot,\cdot)$ of R_l and n_l via

$$f_{R_l,O_l}(r_l,o_l) = f_{R_l,n_l}(r_l,o_l - r_l) = f_{R_l}(r_l)f_{n_l}(o_l - r_l),$$

with the 2nd equality following since R_l & n_l are independent

- the marginal PDF for O_l can be obtained from the joint PDF $f_{R_l,O_l}(\cdot,\cdot)$ by integrating out the first argument:

$$f_{O_l}(o_l) = \int_{-\infty}^{\infty} f_{R_l,O_l}(r_l,o_l)\,dr_l = \int_{-\infty}^{\infty} f_{R_l}(r_l)f_{n_l}(o_l - r_l)\,dr_l$$

- putting all these pieces together yields the conditional PDF

$$f_{R_l|O_l=o_l}(r_l) = \frac{f_{R_l,O_l}(r_l,o_l)}{f_{O_l}(o_l)} = \frac{f_{R_l}(r_l)f_{n_l}(o_l - r_l)}{\int_{-\infty}^{\infty} f_{R_l}(r_l)f_{n_l}(o_l - r_l)\,dr_l}$$

Conditional Mean and Median Approach: III

- mean value of $f_{R_l|O_l=o_l}(\cdot)$ yields estimator $\widehat{R}_l = E\{R_l|O_l\}$:

$$E\{R_l|O_l = o_l\} = \int_{-\infty}^{\infty} r_l f_{R_l|O_l=o_l}(r_l)\, dr_l$$
$$= \frac{\int_{-\infty}^{\infty} r_l f_{R_l}(r_l) f_{n_l}(o_l - r_l) dr_l}{\int_{-\infty}^{\infty} f_{R_l}(r_l) f_{n_l}(o_l - r_l)\, dr_l}$$

- to make further progress, need a model for the transformation-domain representation R_l of the signal
- heuristic that signal in the transformation domain has a few large values and lots of small values suggests a Gaussian mixture model

Conditional Mean and Median Approach: IV

- let $\mathcal{I}_l$ be an RV such that $\mathbf{P}[\mathcal{I}_l = 1] = p_l$ & $\mathbf{P}[\mathcal{I}_l = 0] = 1 - p_l$
- under Gaussian mixture model, R_l has same distribution as

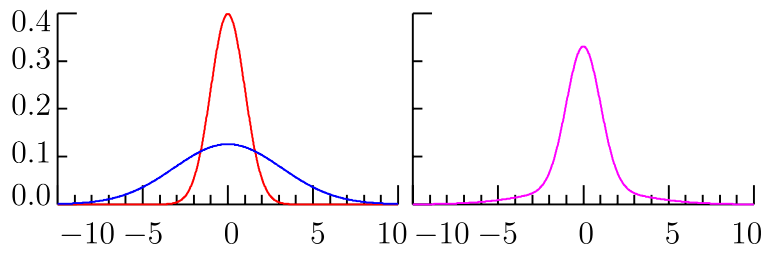
$$\mathcal{I}_l \mathcal{N}(0, \gamma_l^2 \sigma_{G_l}^2) + (1 - \mathcal{I}_l)\mathcal{N}(0, \sigma_{G_l}^2)$$

  where $\mathcal{N}(0, \sigma^2)$ is a Gaussian RV with mean 0 and variance σ^2

  – 2nd component models small # of large signal coefficients

  – 1st component models large # of small coefficients ($\gamma_l^2 \ll 1$)

- example: PDFs for case $\sigma_{G_l}^2 = 10$, $\gamma_l^2 \sigma_{G_l}^2 = 1$ and $p_l = 0.75$



Conditional Mean and Median Approach: V

- to complete model, let n_l obey a Gaussian distribution with mean 0 and variance $\sigma^2_{n_l}$
- conditional mean estimator of the signal RV R_l is given by

$$E\{R_l|O_l = o_l\} = \frac{a_l A_l(o_l) + b_l B_l(o_l)}{A_l(o_l) + B_l(o_l)} o_l,$$

where

$$a_l \equiv \frac{\gamma_l^2 \sigma^2_{G_l}}{\gamma_l^2 \sigma^2_{G_l} + \sigma^2_{n_l}} \text{ and } b_l \equiv \frac{\sigma^2_{G_l}}{\sigma^2_{G_l} + \sigma^2_{n_l}}$$

$$A_l(o_l) \equiv \frac{p_l}{\sqrt{(2\pi[\gamma_l^2 \sigma^2_{G_l} + \sigma^2_{n_l}])}} e^{-o_l^2/[2(\gamma_l^2 \sigma^2_{G_l} + \sigma^2_{n_l})]}$$

$$B_l(o_l) \equiv \frac{1 - p_l}{\sqrt{(2\pi[\sigma^2_{G_l} + \sigma^2_{n_l}])}} e^{-o_l^2/[2(\sigma^2_{G_l} + \sigma^2_{n_l})]}$$

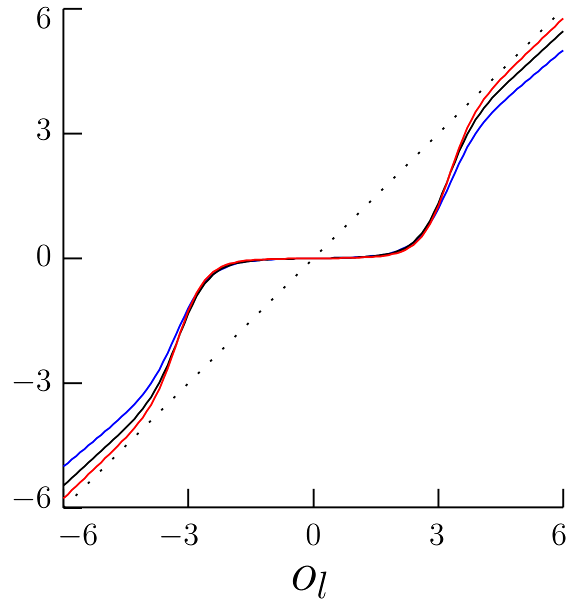
Conditional Mean and Median Approach: VI

- let's simplify to a 'sparse' signal model by setting $\gamma_l = 0$; i.e., large # of small coefficients are all zero
- distribution for R_l same as $(1 - \mathcal{I}_l)\mathcal{N}(0, \sigma^2_{G_l})$
- conditional mean estimator becomes $E\{R_l | O_l = o_l\} = \frac{b_l}{1+c_l} o_l$, where

$$c_l = \frac{p_l \sqrt{(\sigma^2_{G_l} + \sigma^2_{n_l})}}{(1 - p_l)\sigma_{n_l}} e^{-o_l^2 b_l/(2\sigma^2_{n_l})}$$

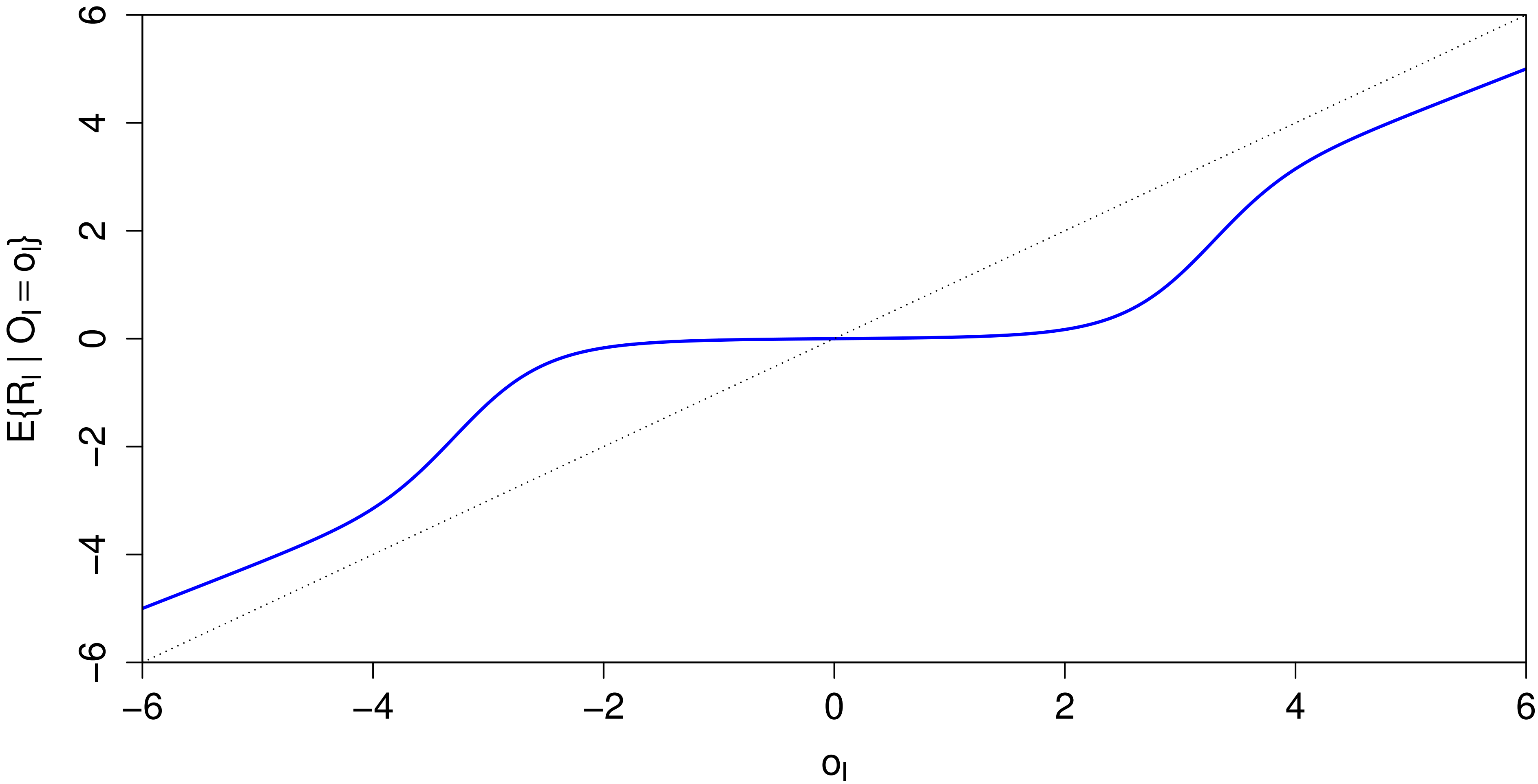
Conditional Mean and Median Approach: VII

$E\{R_l | O_l = o_l\}$

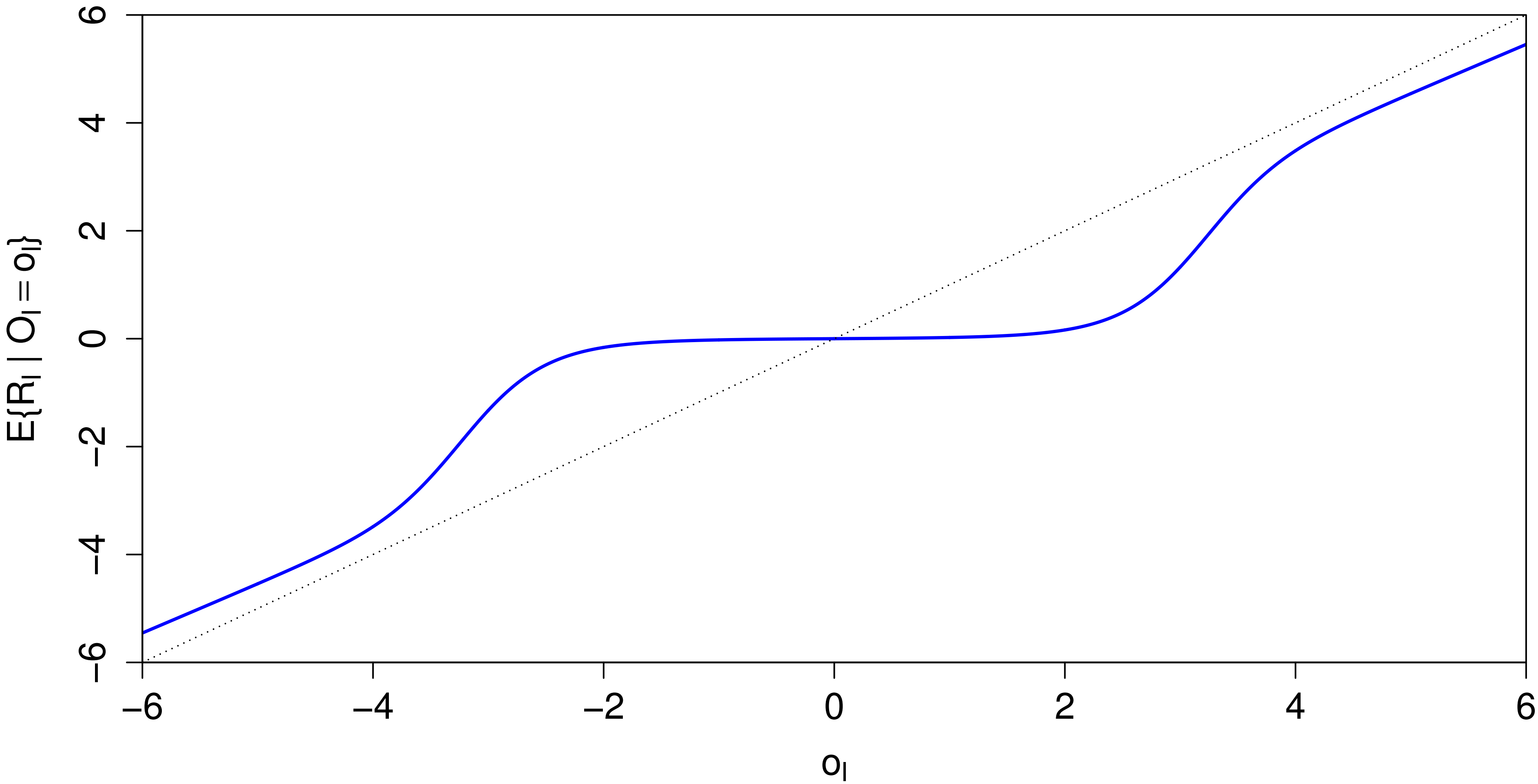


- conditional mean shrinkage rule for $p_l = 0.95$ (i.e., $\approx$ 95% of signal coefficients are 0); $\sigma^2_{n_l} = 1$; and $\sigma^2_{G_l} = 5$ (curve furthest from dotted diagonal), 10 and 25 (curve nearest to diagonal)
- as $\sigma^2_{G_l}$ gets large (i.e., large signal coefficients increase in size), shrinkage rule starts to resemble mid thresholding rule

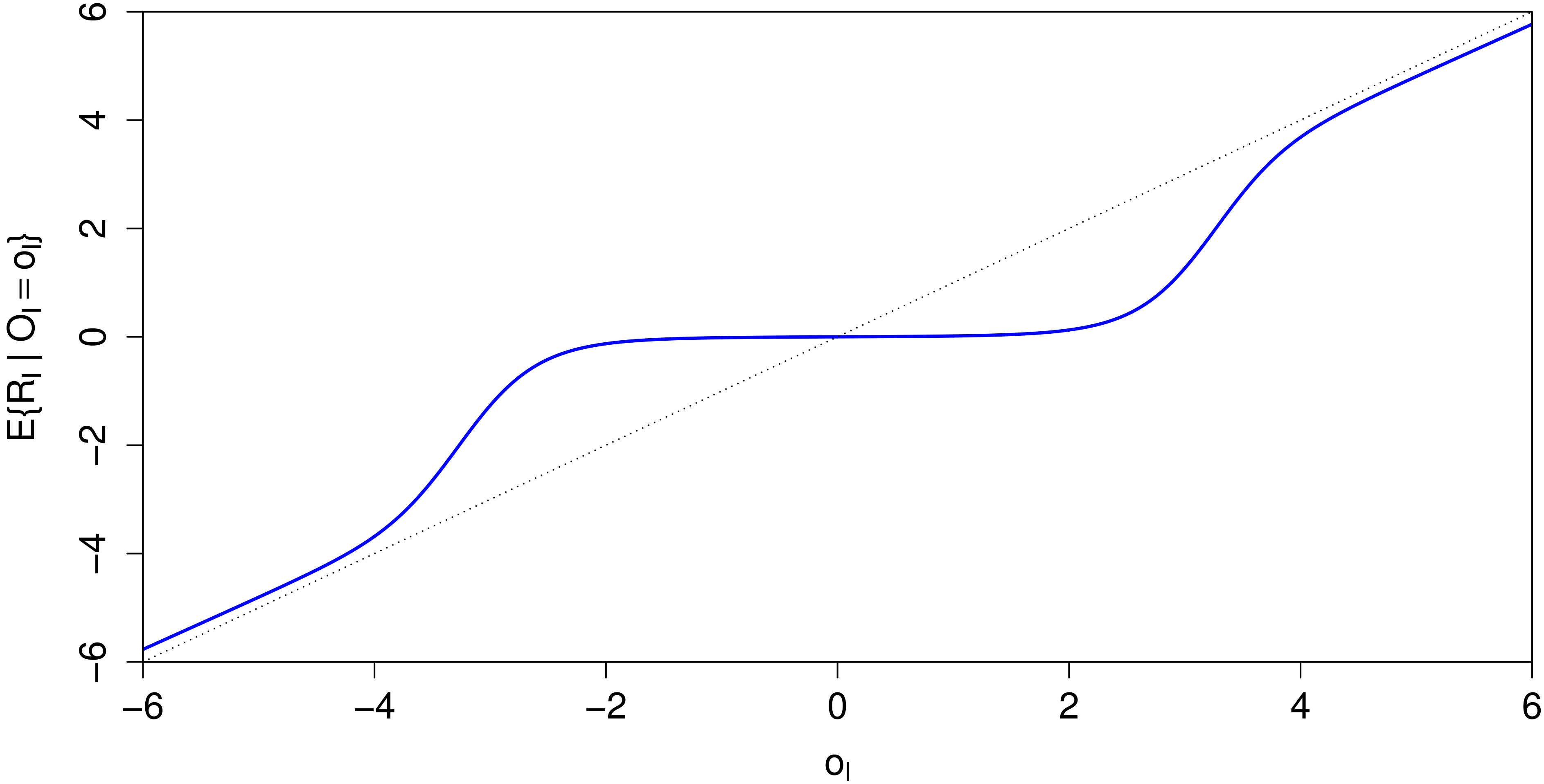
$p_l = 0.95$, $\sigma^2_{n_l} = 1$ **and** $\sigma^2_{G_l} = 5$



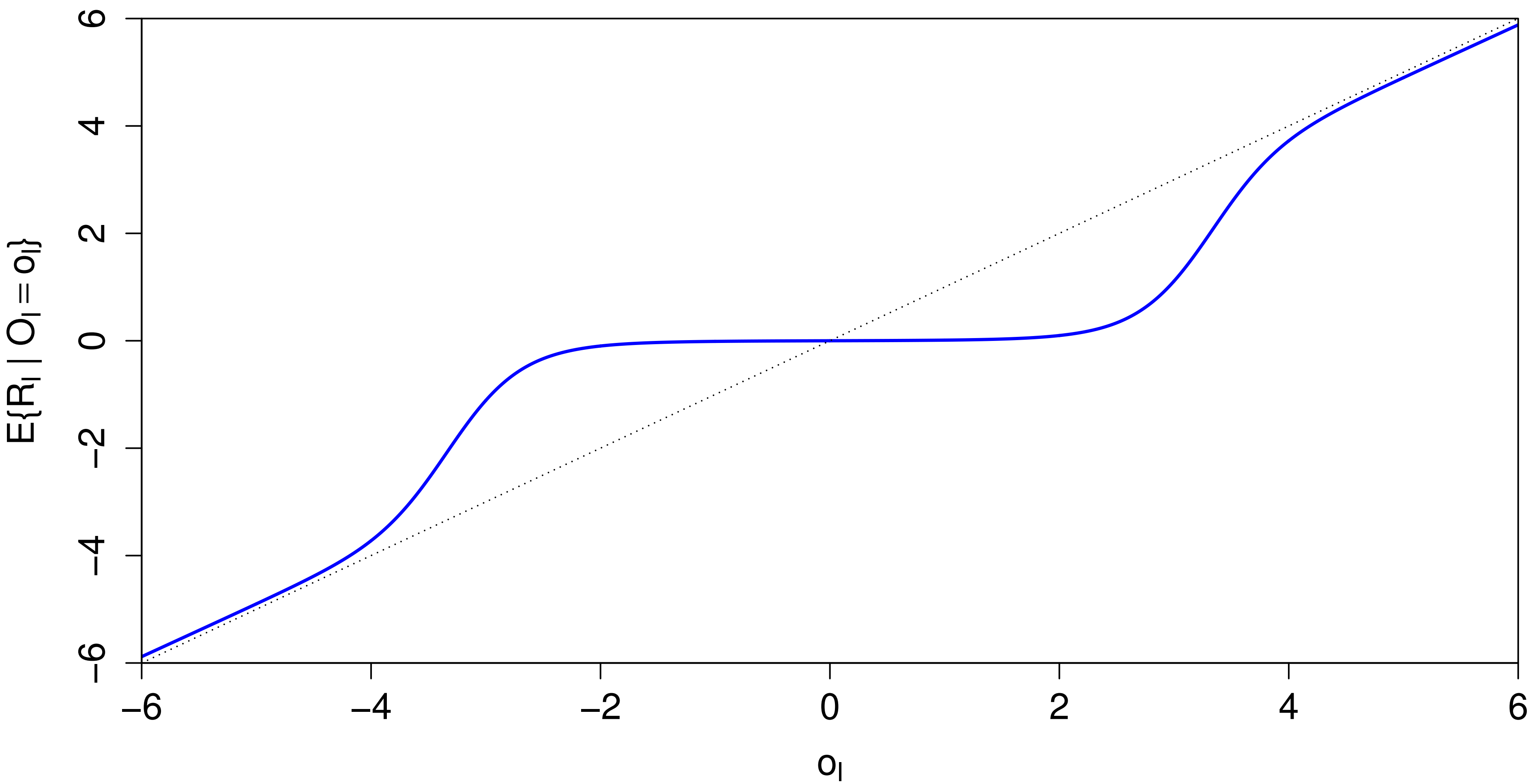
$p_l = 0.95$, $\sigma^2_{n_l} = 1$ **and** $\sigma^2_{G_l} = 10$



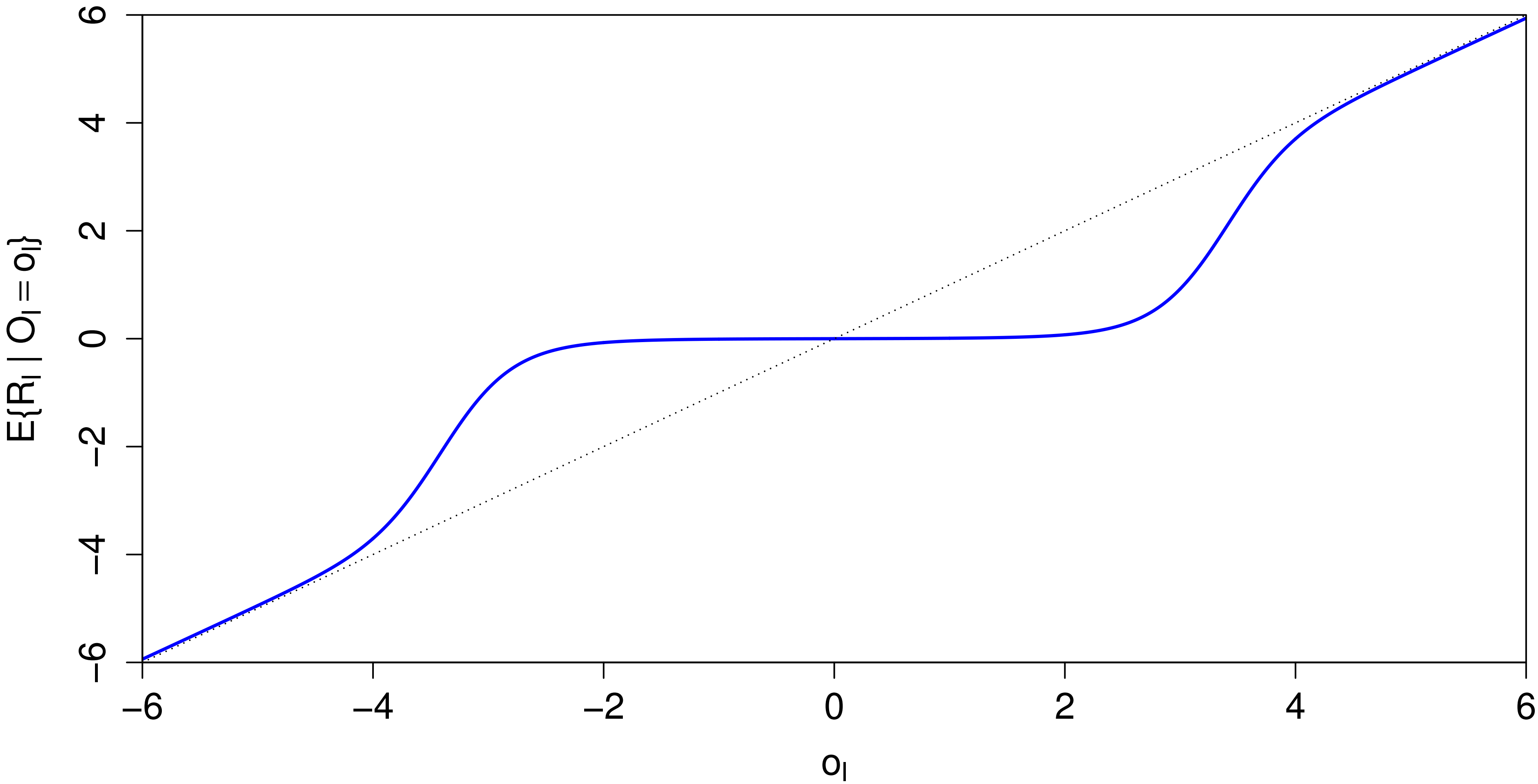
$p_l = 0.95$, $\sigma^2_{n_l} = 1$ and $\sigma^2_{G_l} = 25$



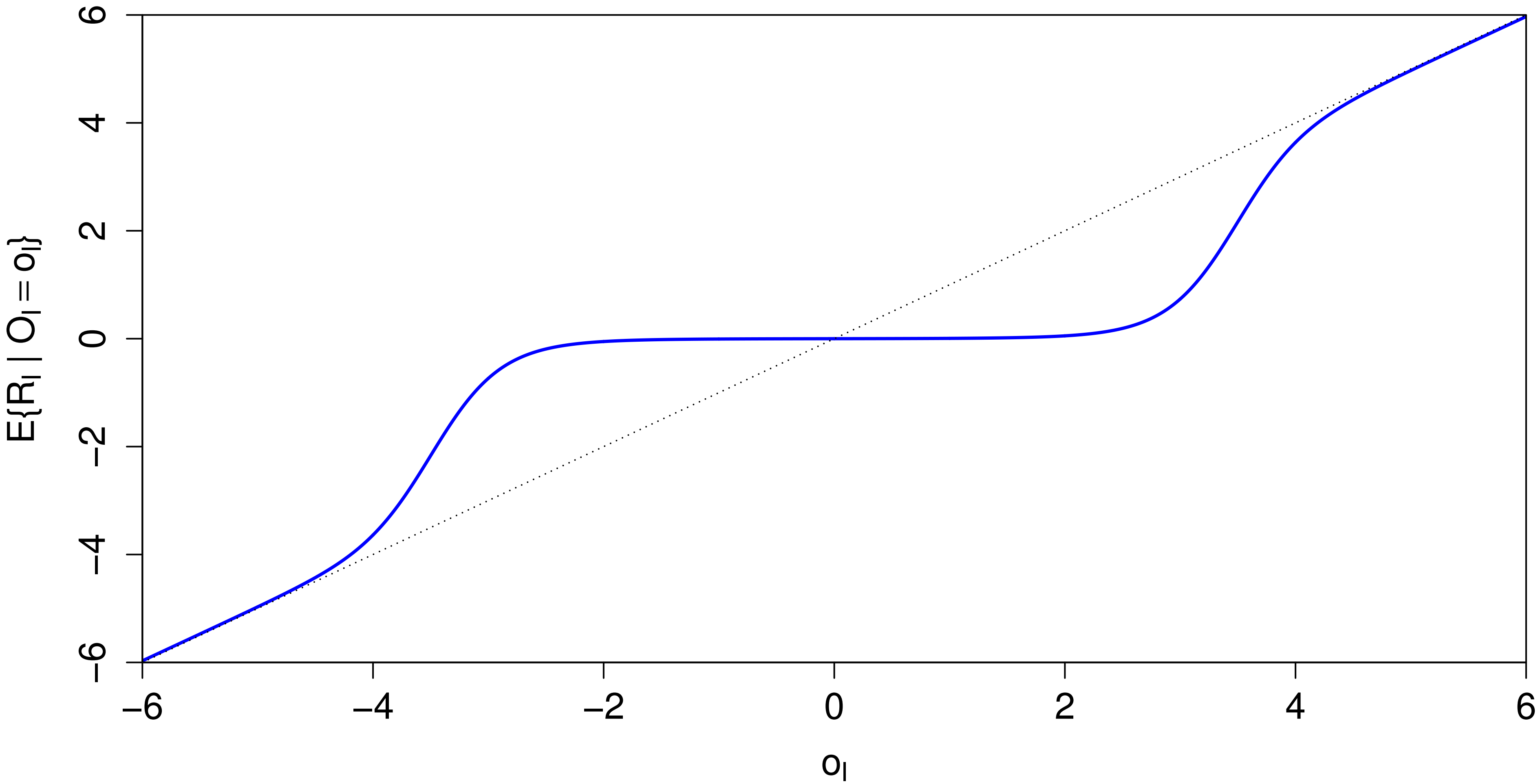
$p_l = 0.95$**,** $\sigma^2_{n_l} = 1$ **and** $\sigma^2_{G_l} = 50$



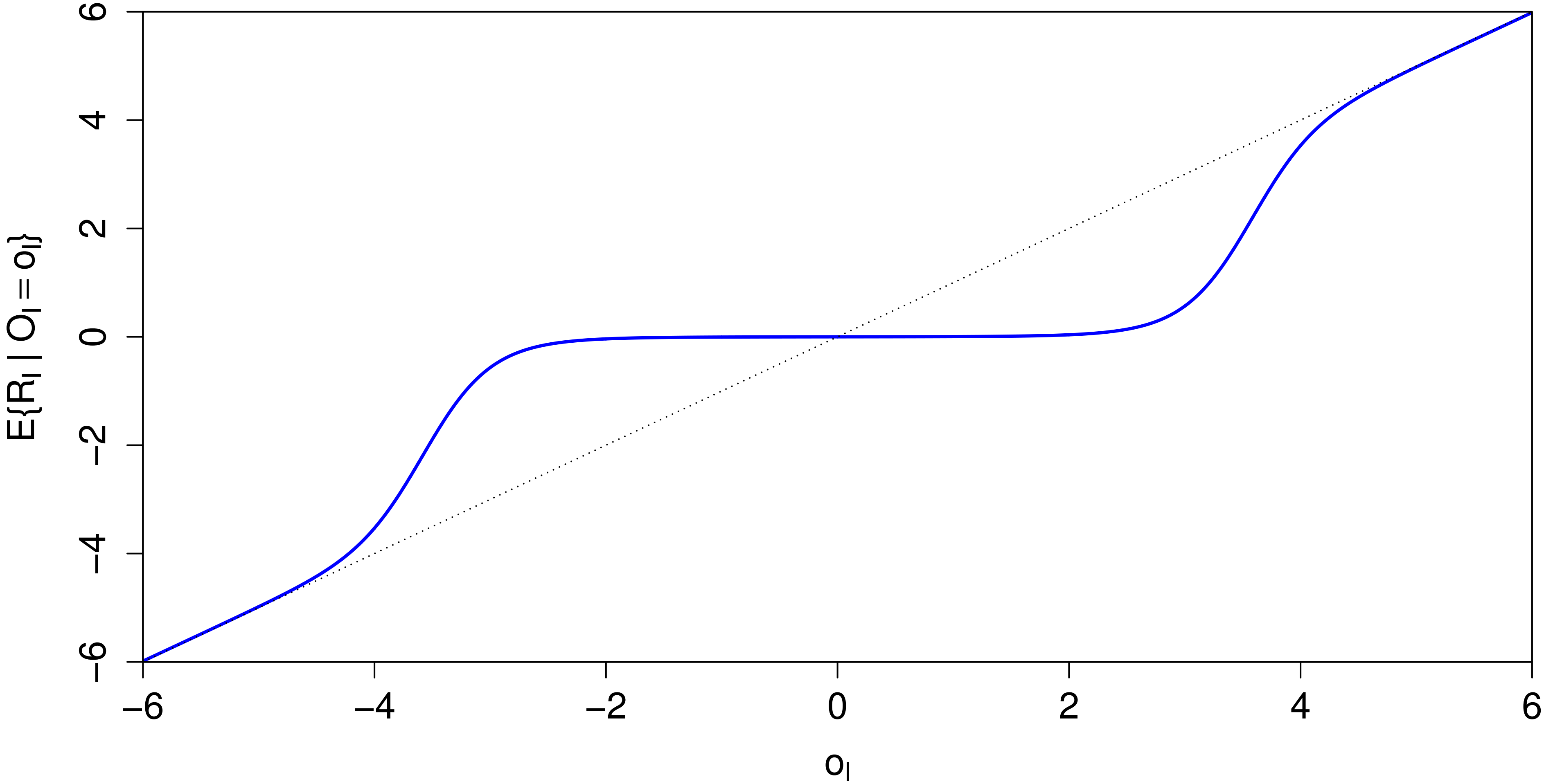
$p_l = 0.95$**,** $\sigma^2_{n_l} = 1$ **and** $\sigma^2_{G_l} = 100$



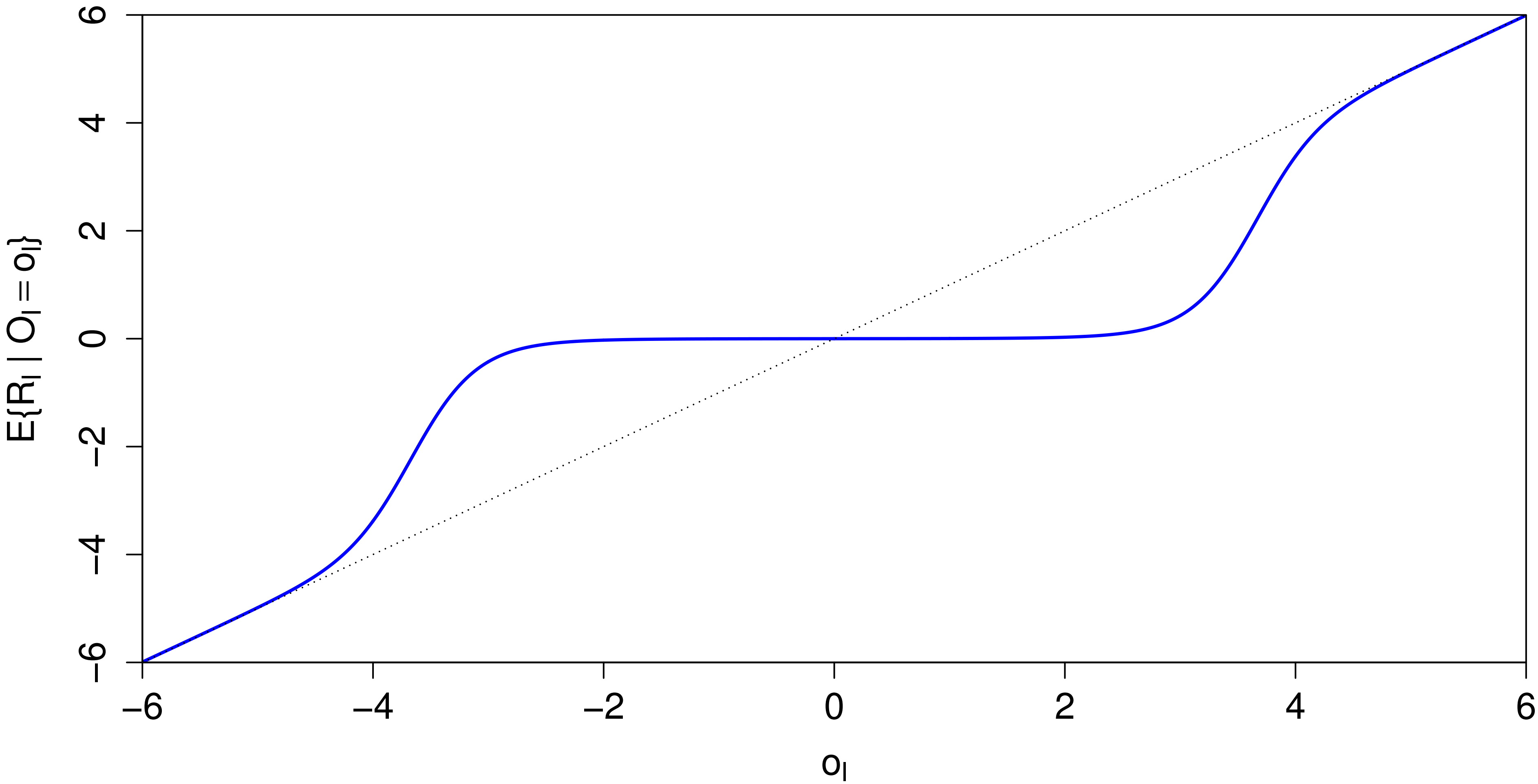
$p_l = 0.95$**,** $\sigma^2_{n_l} = 1$ **and** $\sigma^2_{G_l} = 200$



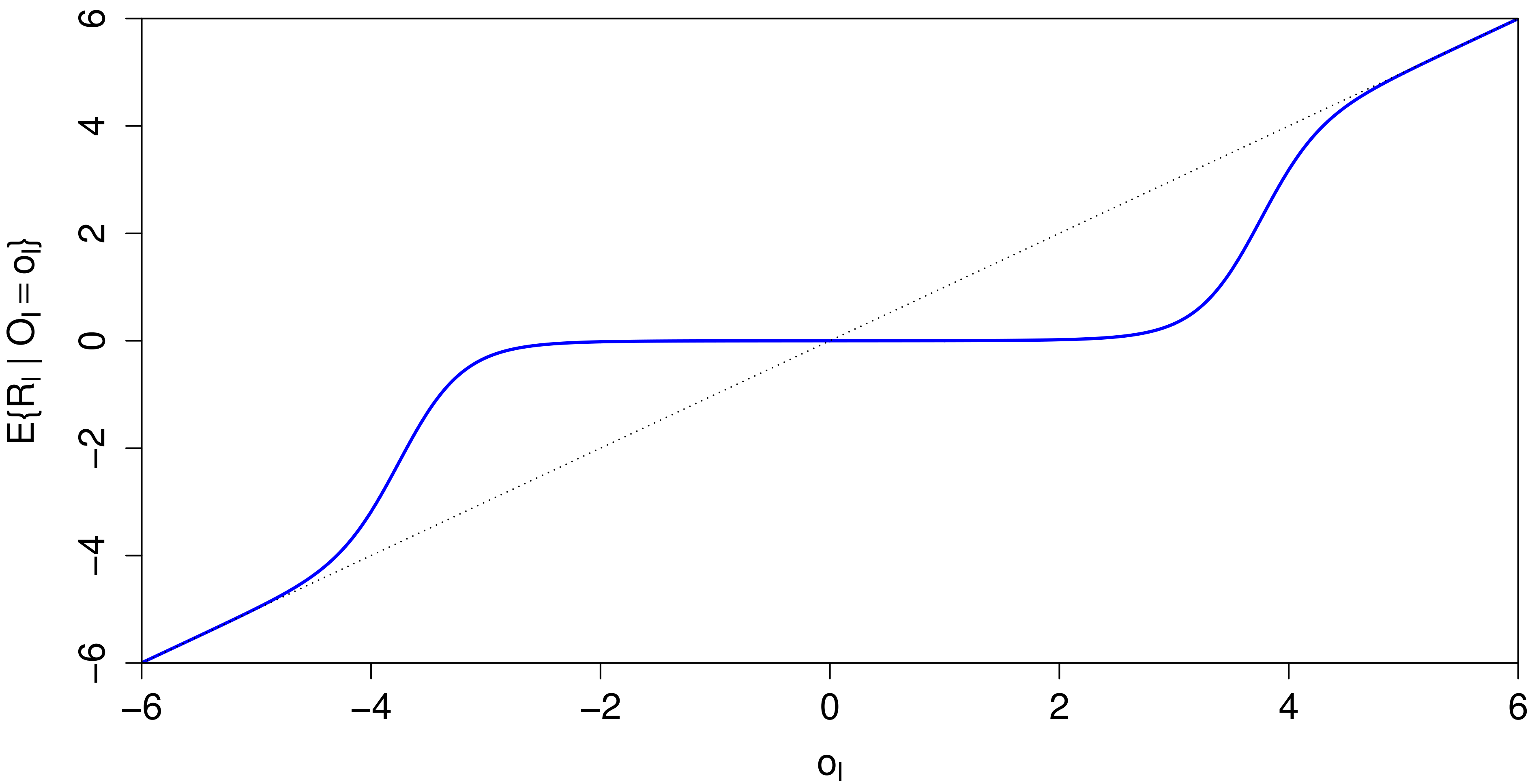
$p_l = 0.95$, $\sigma^2_{n_l} = 1$ **and** $\sigma^2_{G_l} = 400$



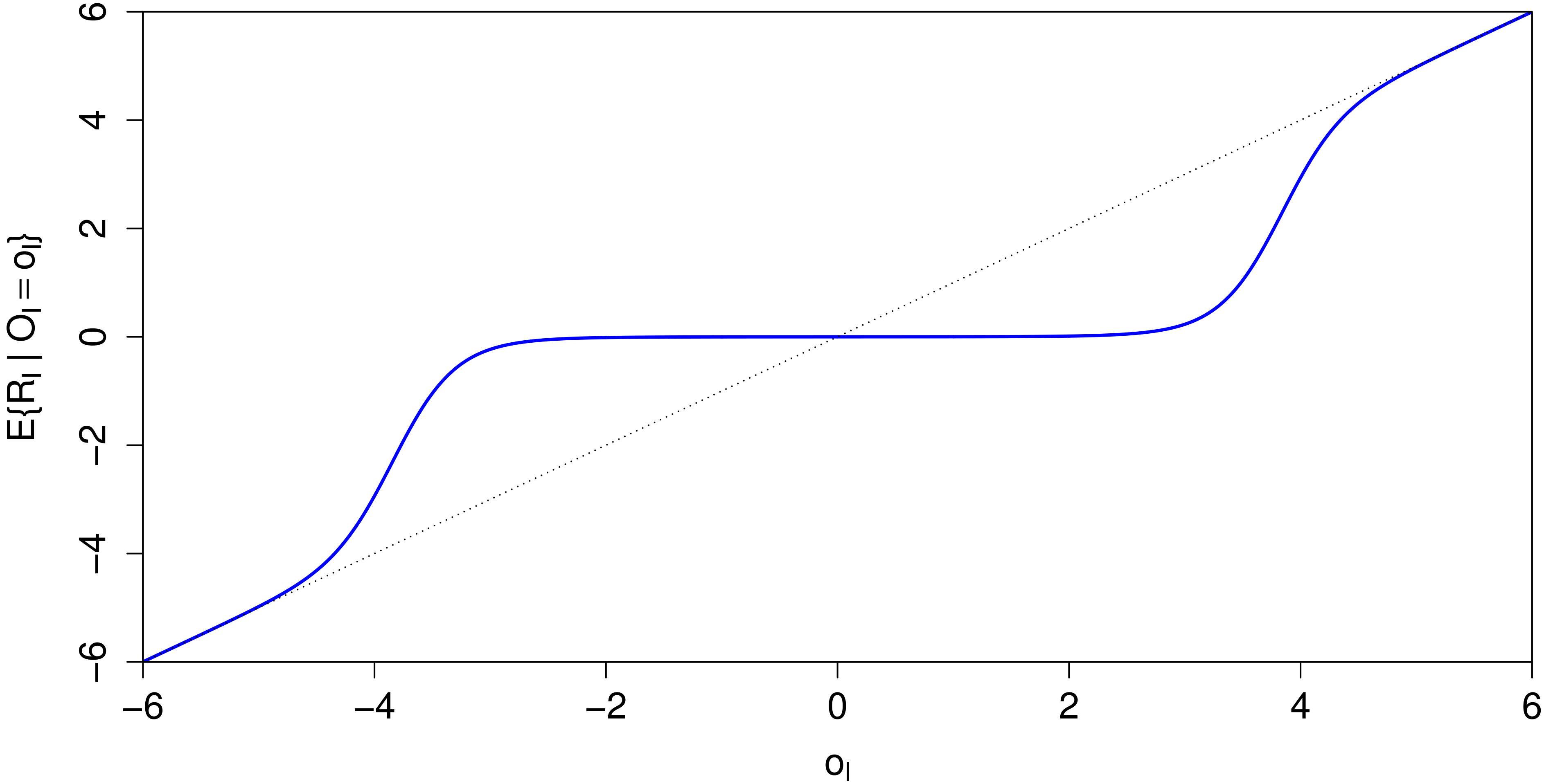
$p_l = 0.95$, $\sigma^2_{n_l} = 1$ **and** $\sigma^2_{G_l} = 800$



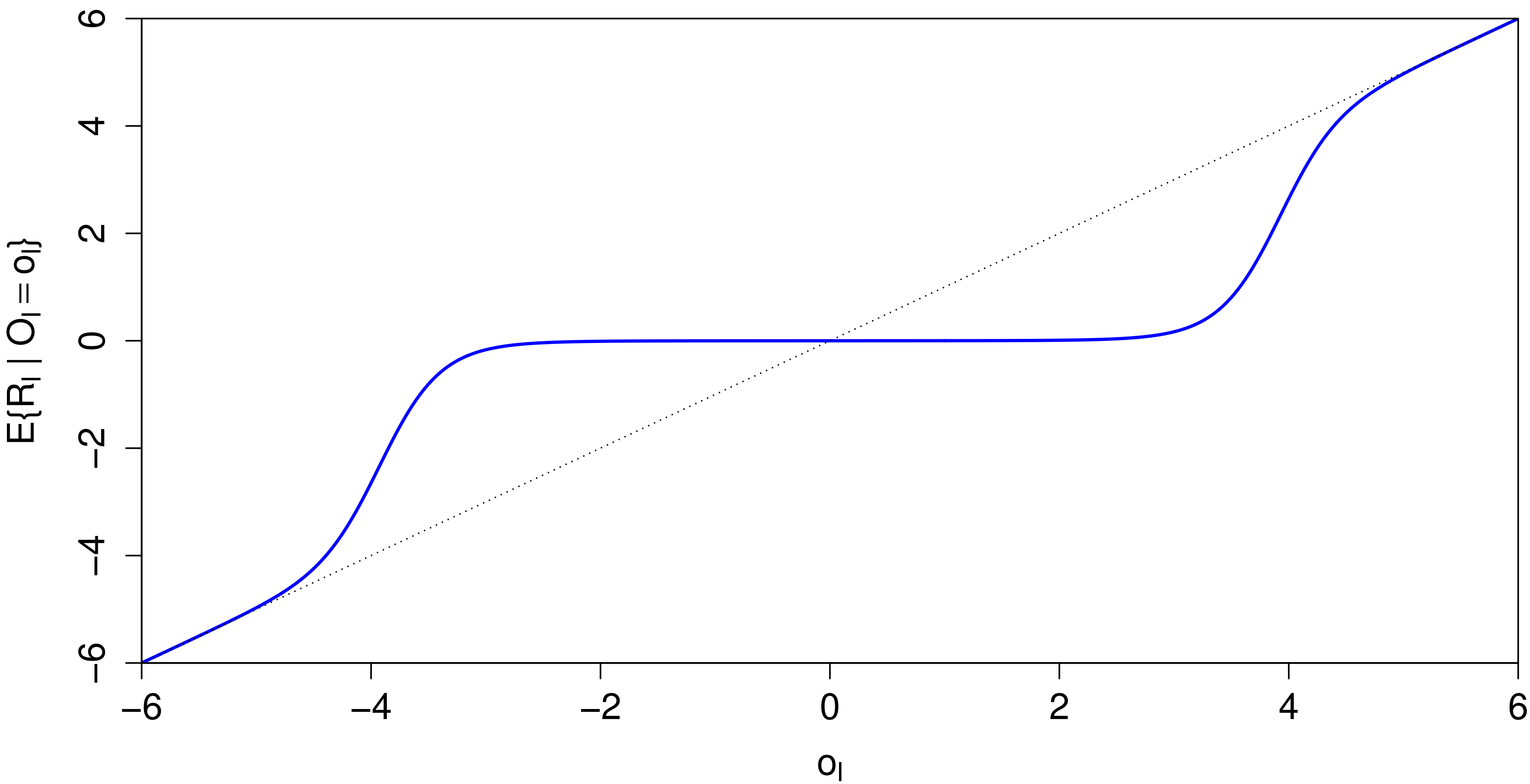
$p_l = 0.95$, $\sigma^2_{n_l} = 1$ **and** $\sigma^2_{G_l} = 1600$



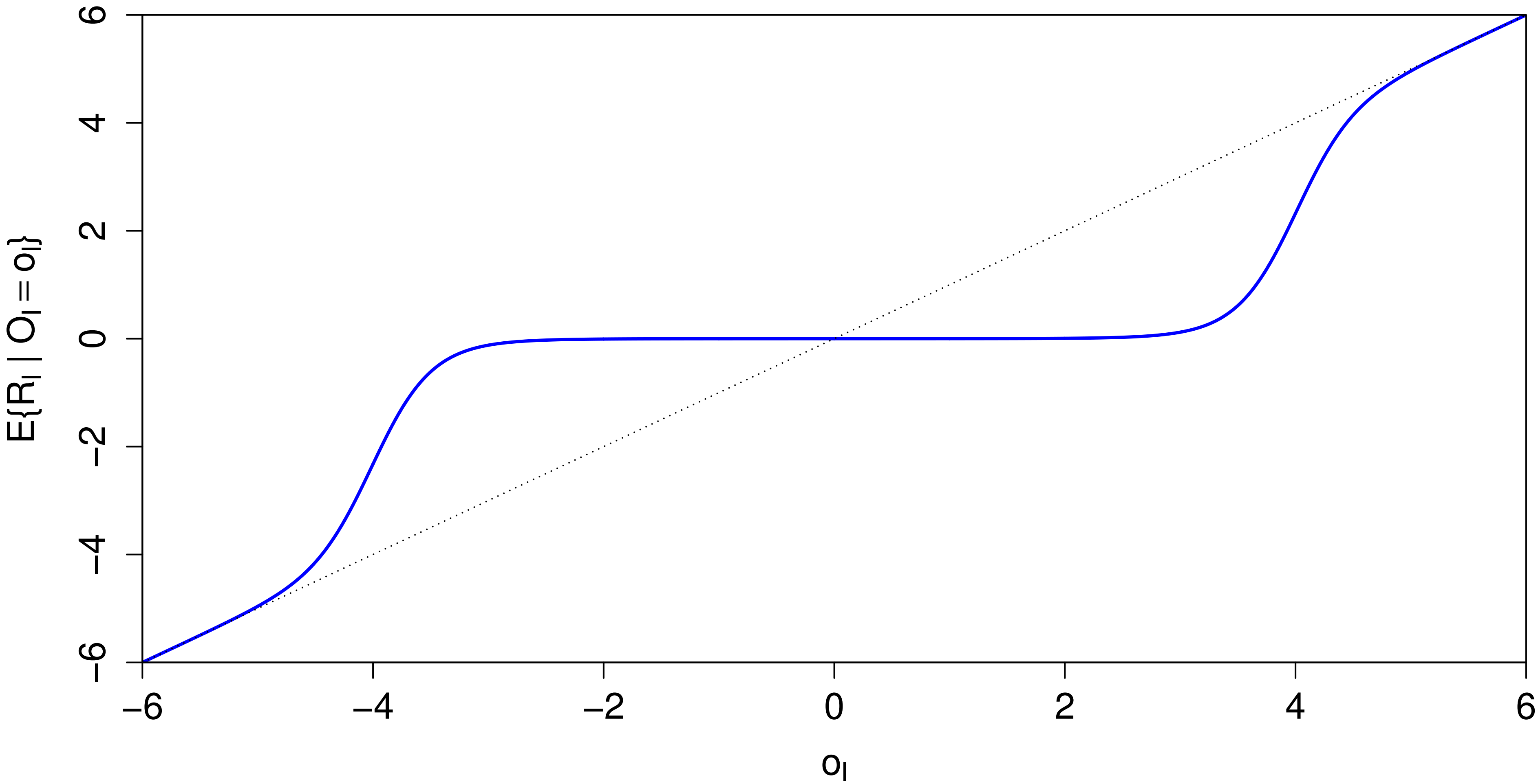
$p_l = 0.95$, $\sigma^2_{n_l} = 1$ **and** $\sigma^2_{G_l} = 3200$



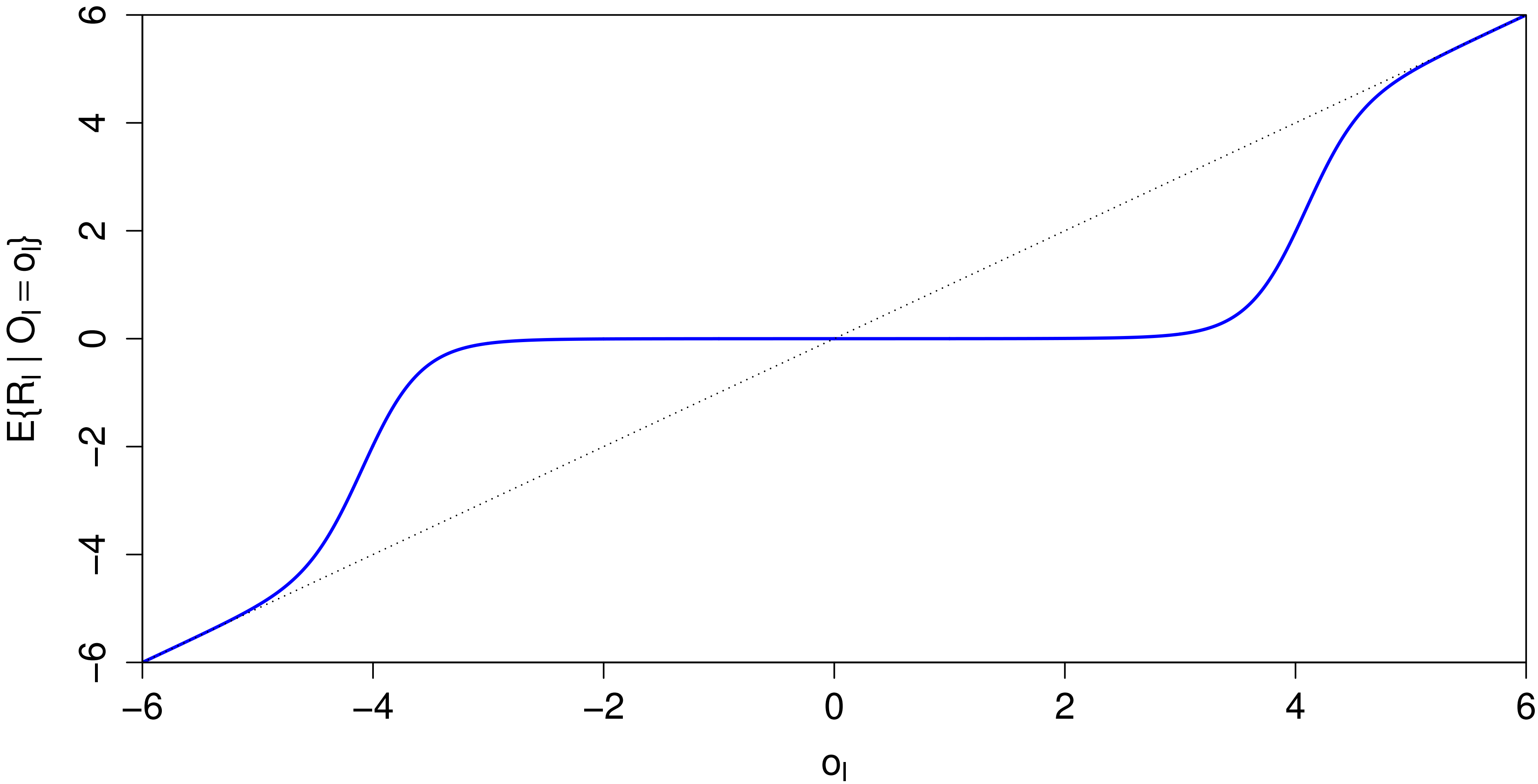
$p_l = 0.95$**,** $\sigma^2_{n_l} = 1$ **and** $\sigma^2_{G_l} = 6400$



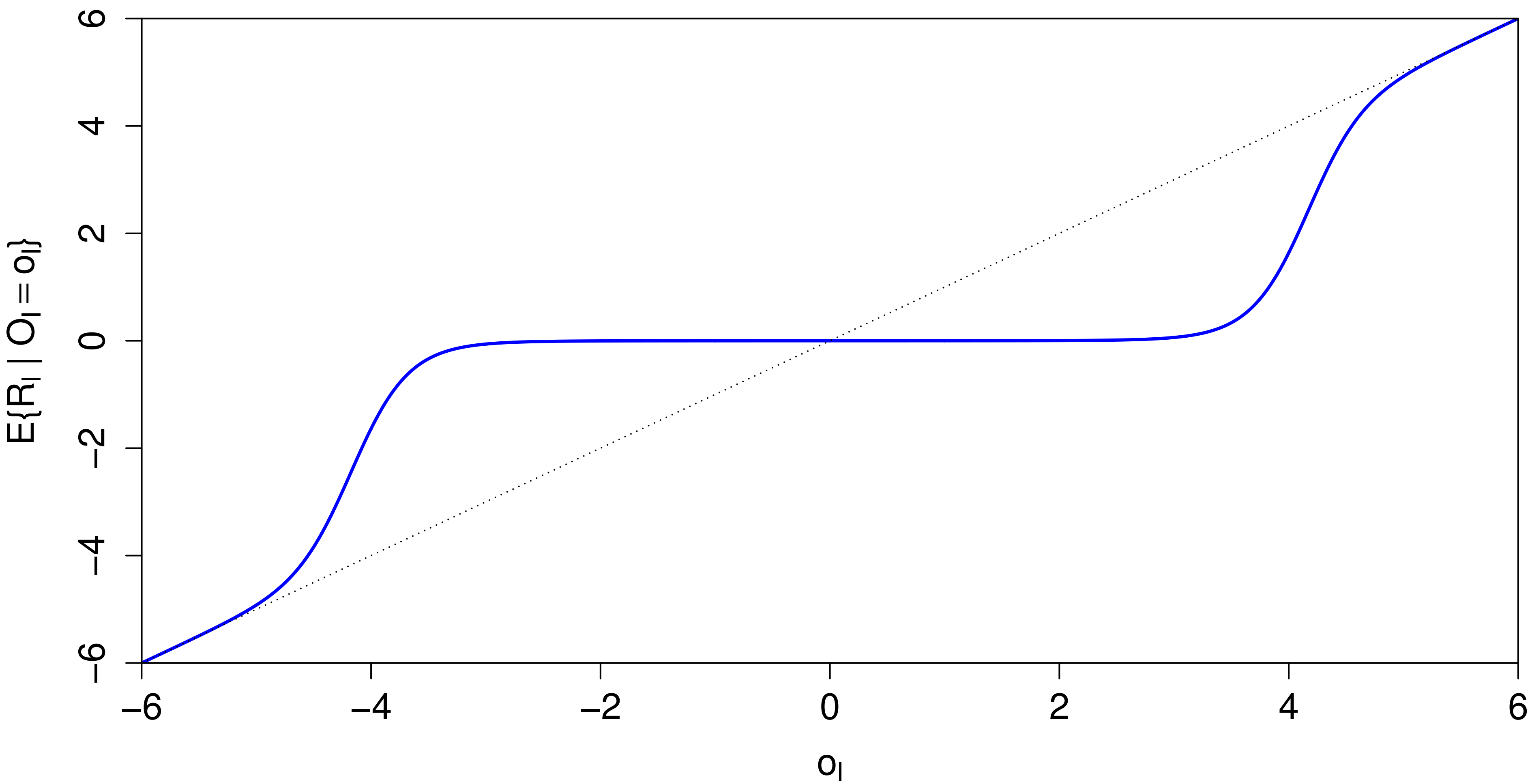
$p_l = 0.95$, $\sigma^2_{n_l} = 1$ **and** $\sigma^2_{G_l} = 12800$



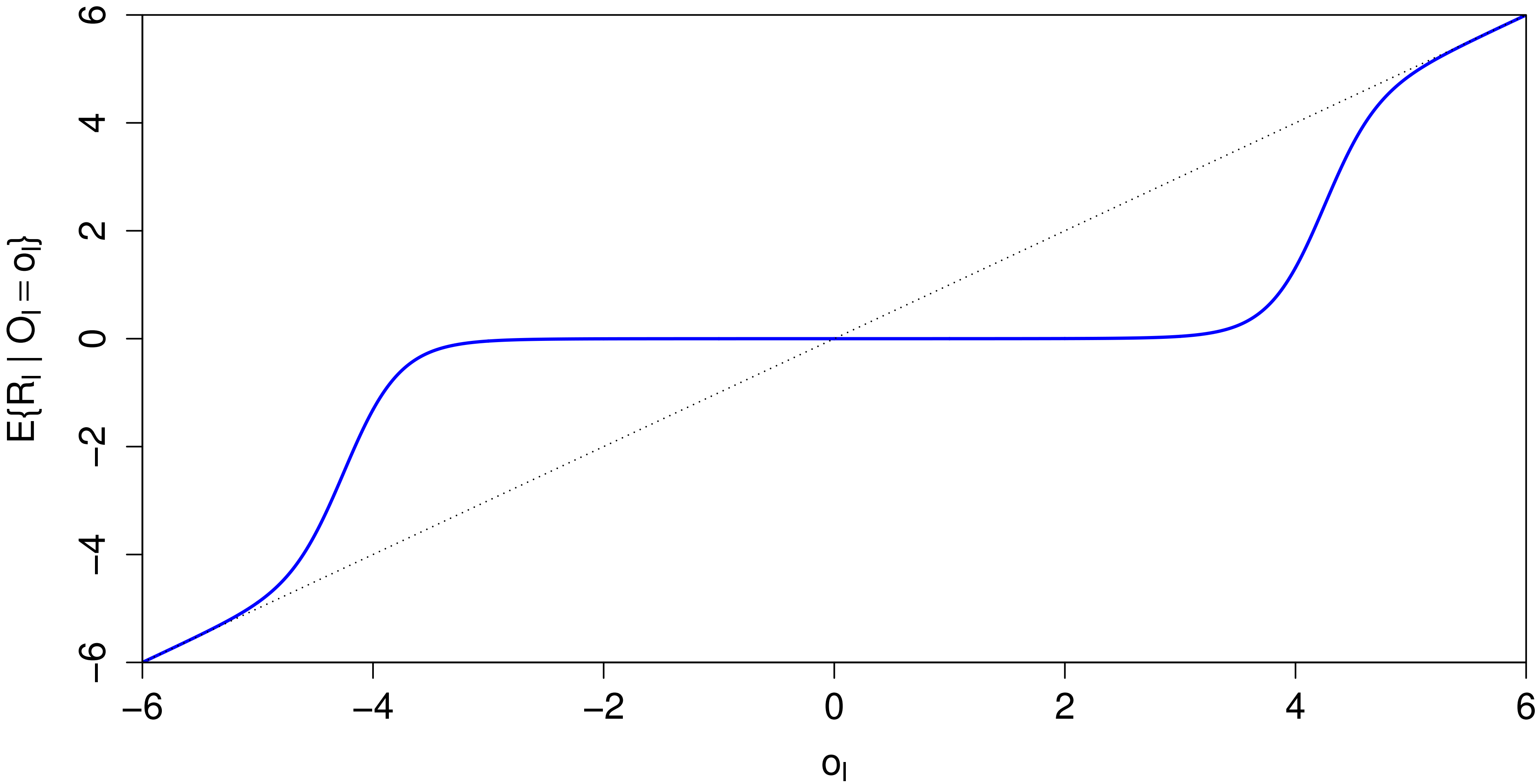
$p_l = 0.95$**,** $\sigma^2_{n_l} = 1$ **and** $\sigma^2_{G_l} = 25600$



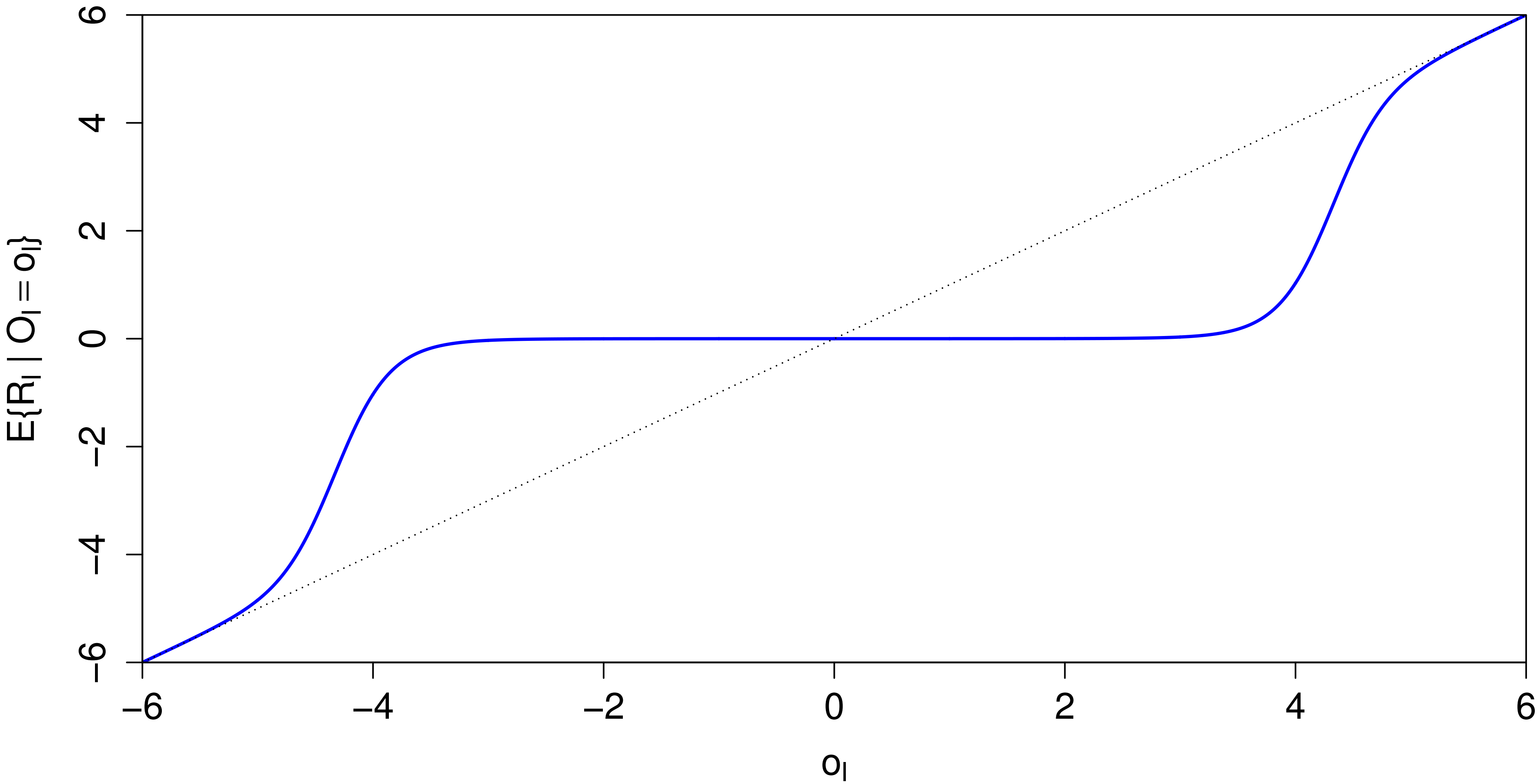
$$p_l = 0.95, \ \sigma^2_{n_l} = 1 \ \textbf{and} \ \sigma^2_{G_l} = 51200$$



$p_l = 0.95$, $\sigma^2_{n_l} = 1$ **and** $\sigma^2_{G_l} = 102400$



$p_l = 0.95$, $\sigma^2_{n_l} = 1$ **and** $\sigma^2_{G_l} = 204800$



Conditional Mean and Median Approach: VIII

- now suppose we estimate R_l via $\widehat{R}_l = U_1(O_l)$, where $U_1(O_l)$ is selected to minimize $E\{|R_l - U_1(O_l)|\}$
- solution is to set $U_1(o_l)$ to the median of the PDF for R_l given $O_l = o_l$
- to find $U_1(o_l)$, need to solve for it in the equation

$$\int_{-\infty}^{U_1(o_l)} f_{R_l|O_l=o_l}(r_l)\, dr_l = \frac{\int_{-\infty}^{U_1(o_l)} f_{R_l}(r_l) f_{n_l}(o_l - r_l)\, dr_l}{\int_{-\infty}^{\infty} f_{R_l}(r_l) f_{n_l}(o_l - r_l)\, dr_l} = \frac{1}{2}$$

Conditional Mean and Median Approach: IX

- simplifying to the sparse signal model, Godfrey & Rocca (1981) show that

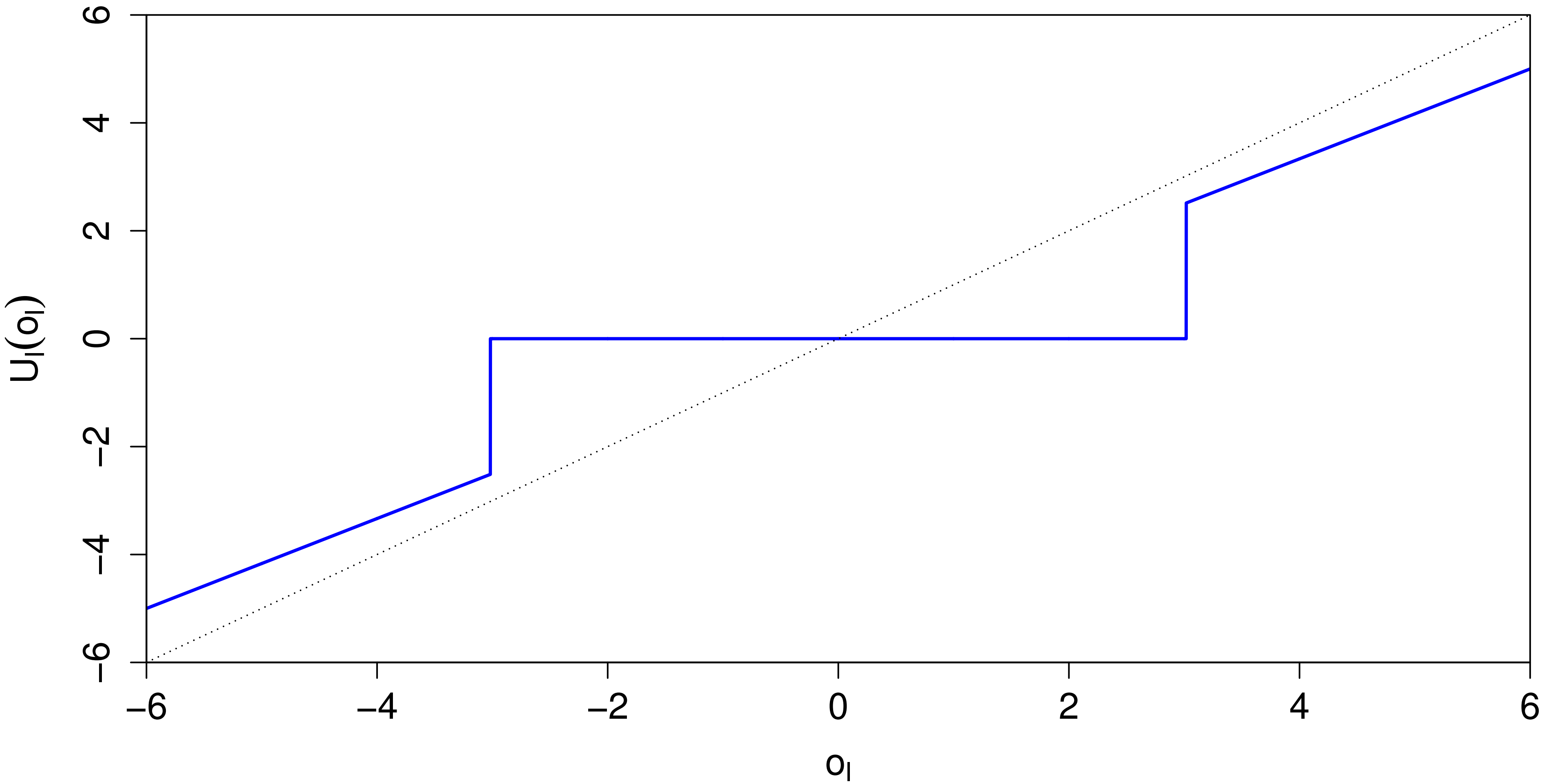
$$U_1(O_l) \approx \begin{cases} 0, & \text{if } |O_l| \leq \delta; \\ b_l O_l, & \text{otherwise,} \end{cases}$$

where

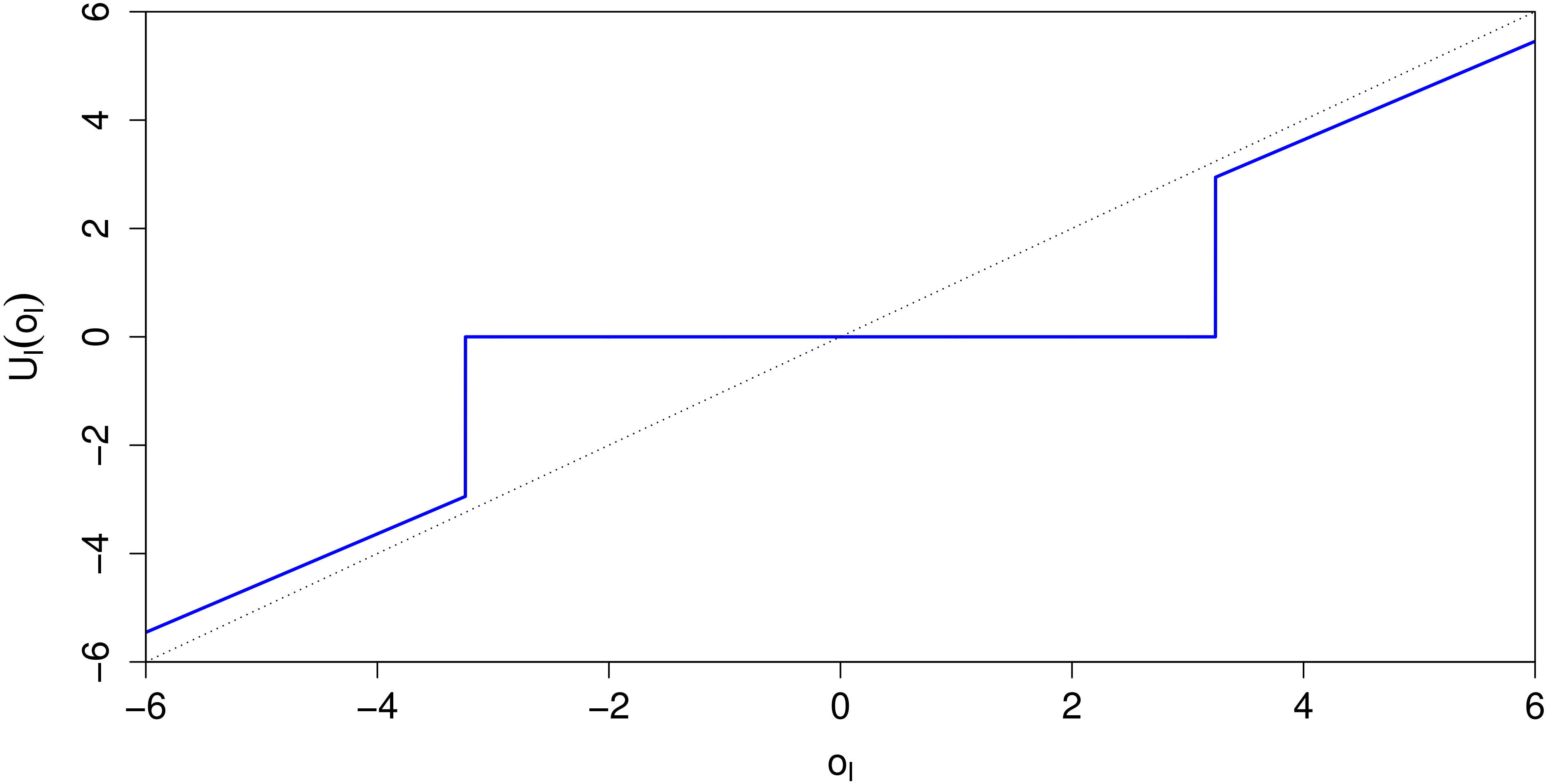
$$\delta = \sigma_{n_l} \left[2 \log \left(\frac{p_l \sigma_{G_l}}{(1 - p_l)\sigma_{n_l}} \right) \right]^{1/2} \quad \text{and} \quad b_l = \frac{\sigma^2_{G_l}}{\sigma^2_{G_l} + \sigma^2_{n_l}}$$

- above approximation valid if $p_l/(1 - p_l) \gg \sigma^2_{n_l}/(\sigma_{G_l}\delta)$ and $\sigma^2_{G_l} \gg \sigma^2_{n_l}$
- note that $U_1(\cdot)$ is approximately a hard thresholding rule

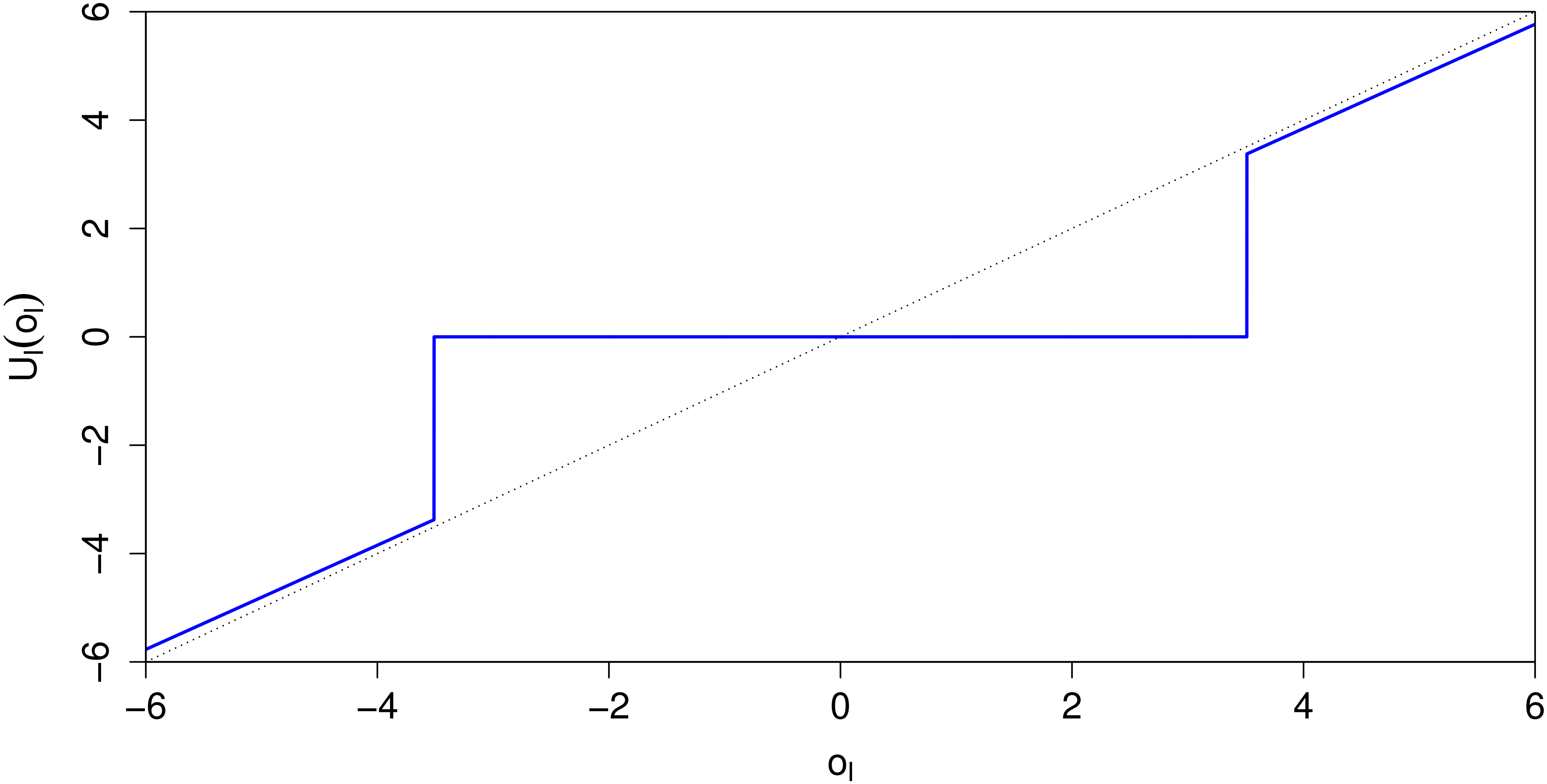
$p_l = 0.95$, $\sigma^2_{n_l} = 1$ **and** $\sigma^2_{G_l} = 5$



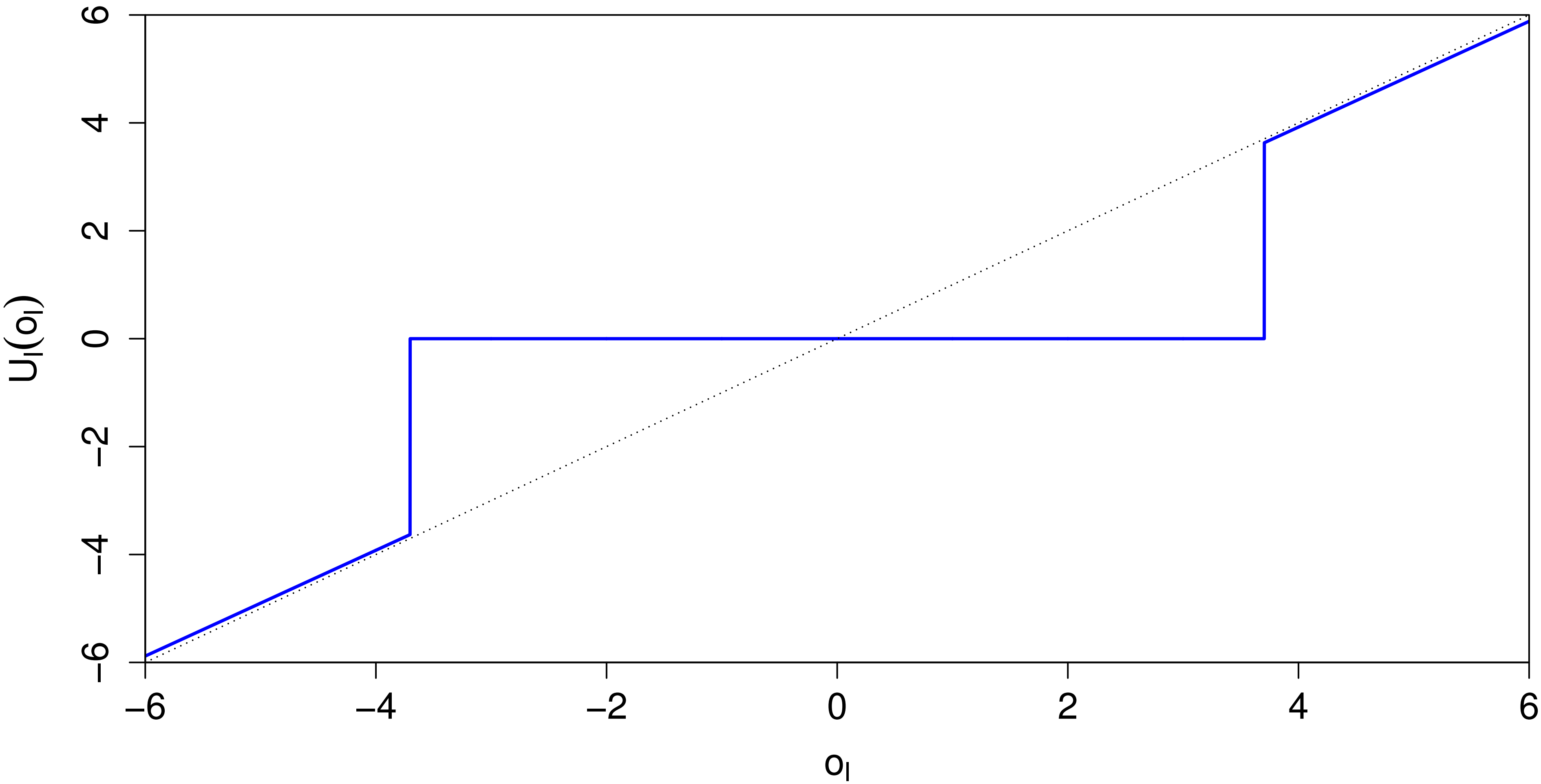
$p_l = 0.95$, $\sigma^2_{n_l} = 1$ **and** $\sigma^2_{G_l} = 10$



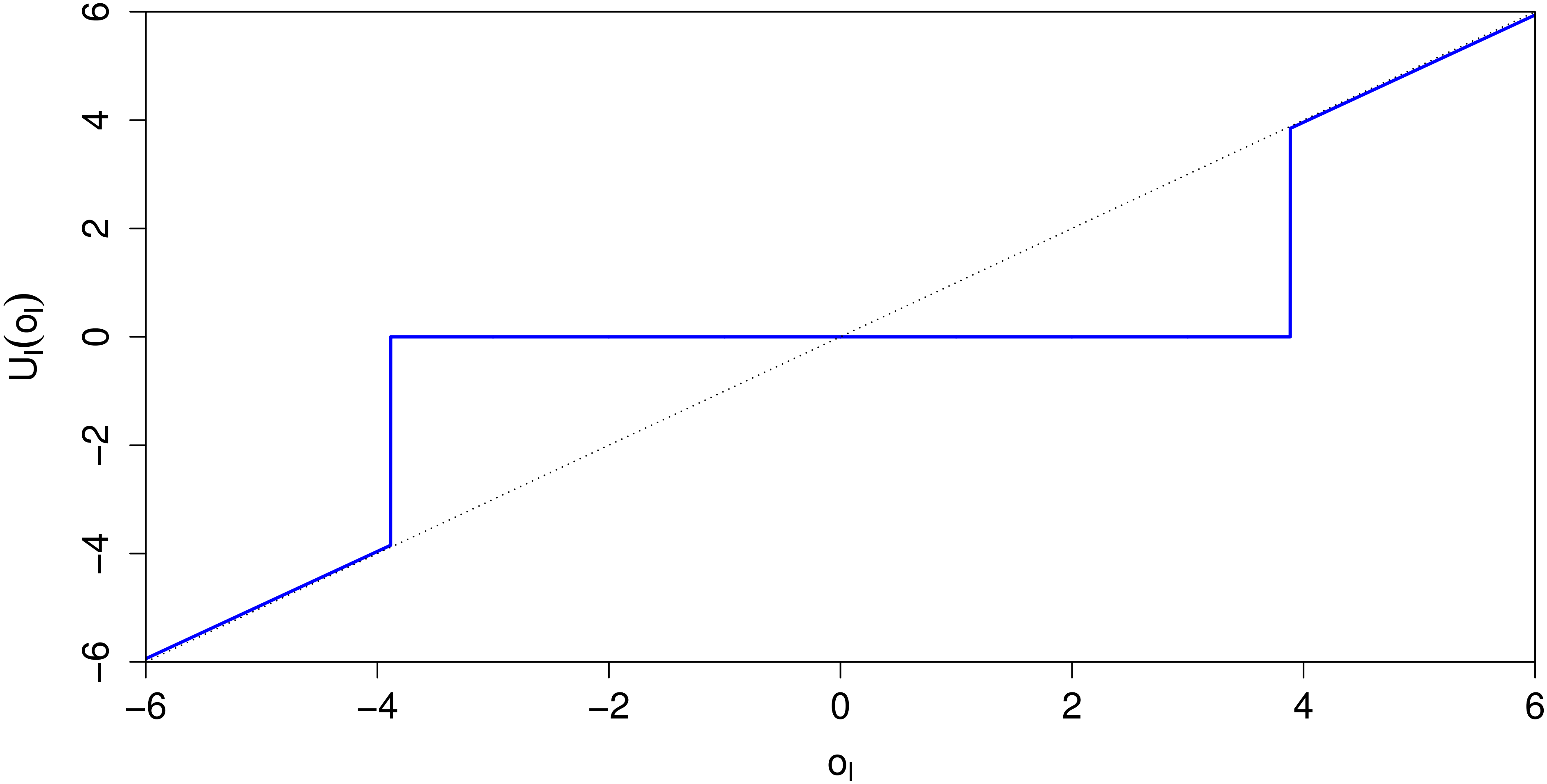
$p_l = 0.95$, $\sigma^2_{n_l} = 1$ **and** $\sigma^2_{G_l} = 25$



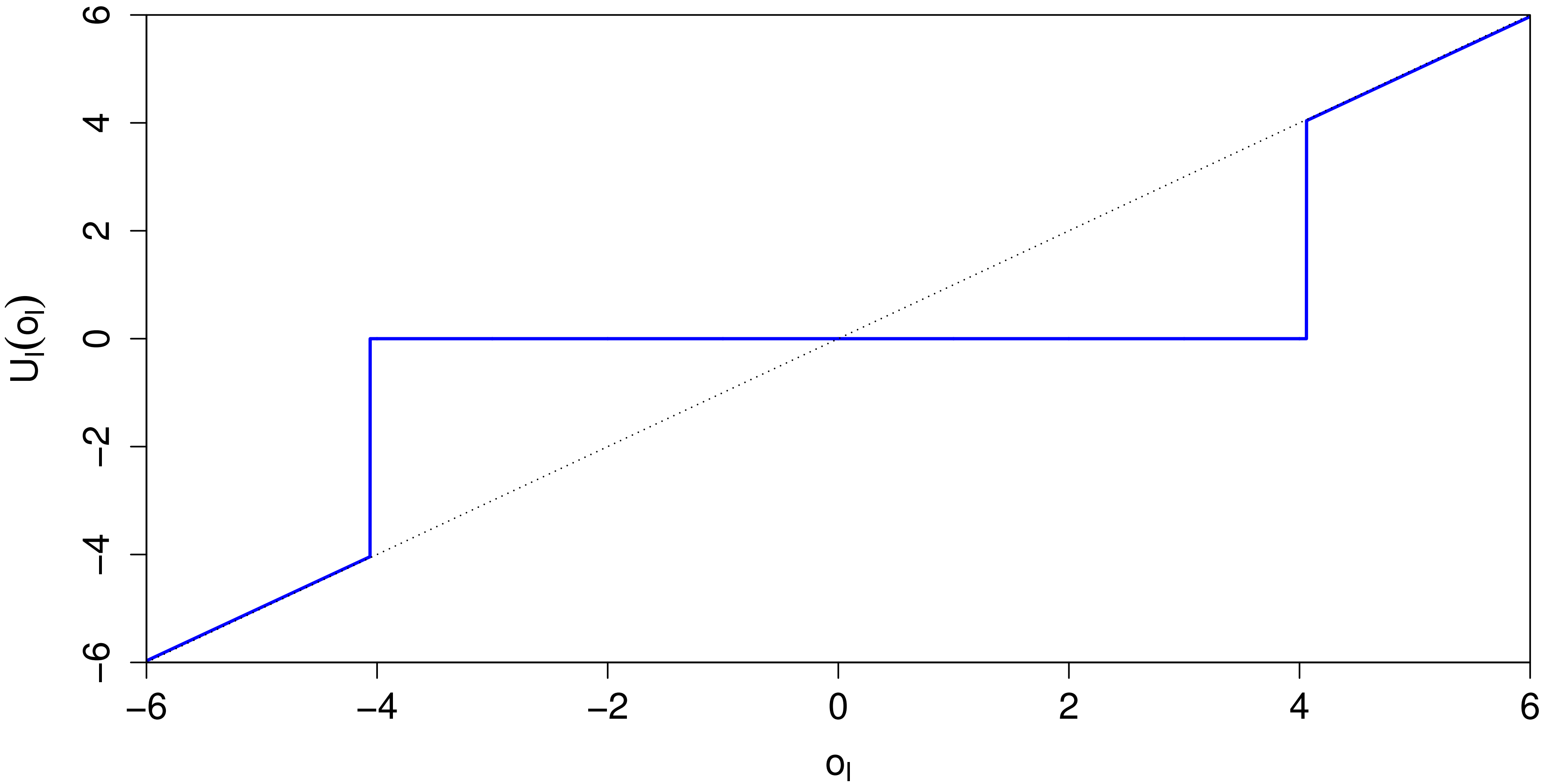
$p_l = 0.95$, $\sigma^2_{n_l} = 1$ **and** $\sigma^2_{G_l} = 50$



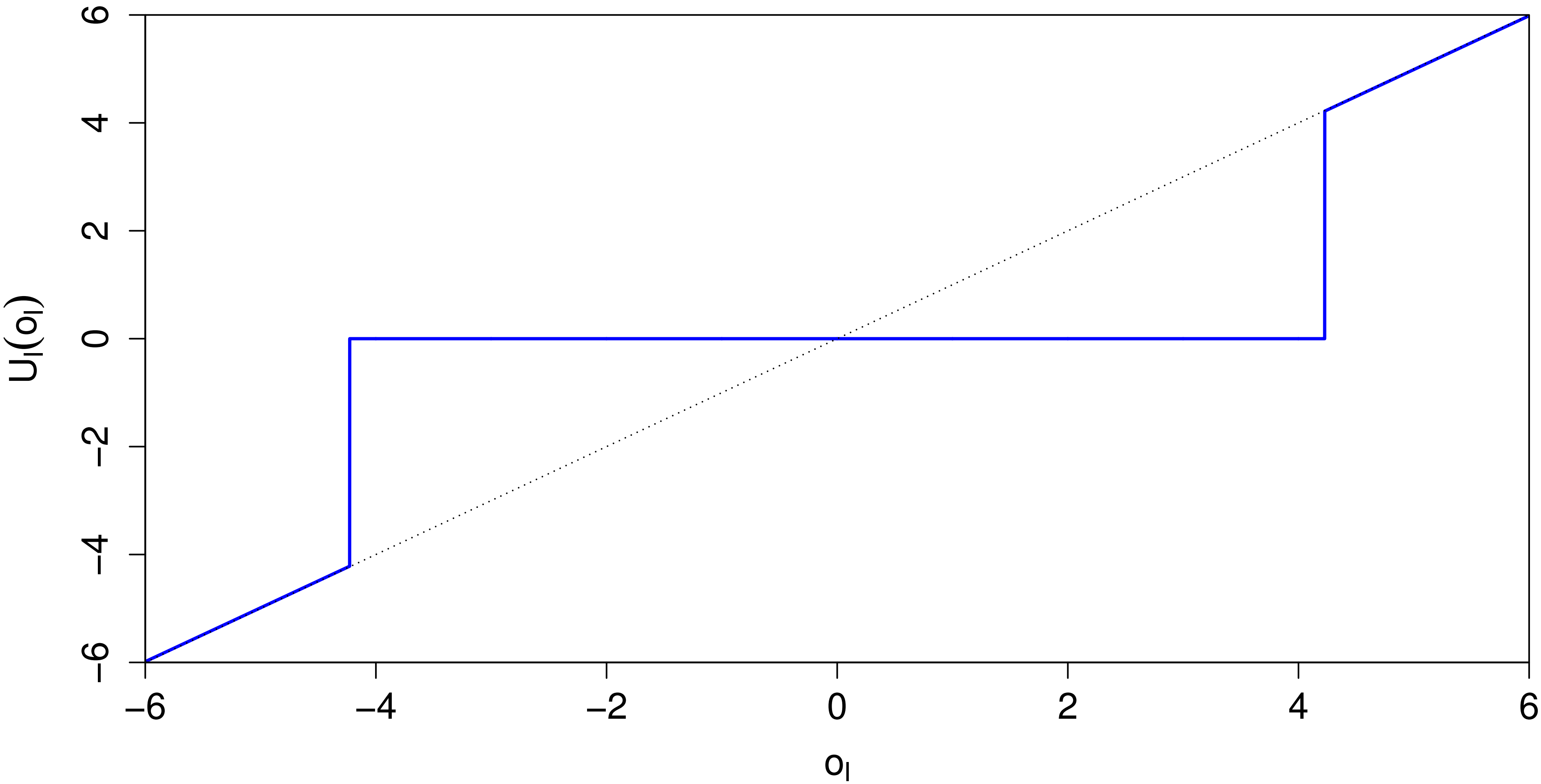
$p_l = 0.95$, $\sigma^2_{n_l} = 1$ **and** $\sigma^2_{G_l} = 100$



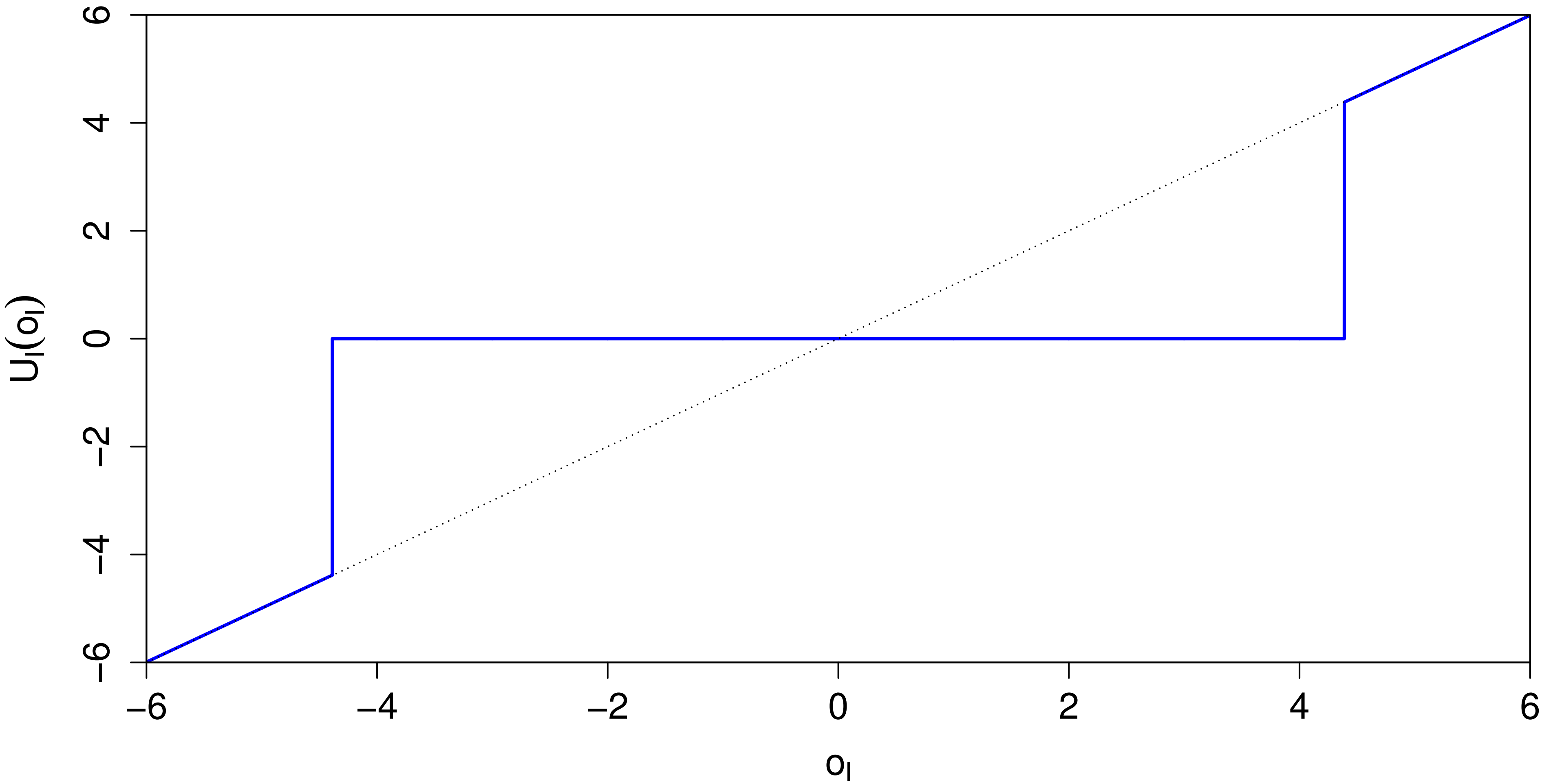
$p_l = 0.95$, $\sigma^2_{n_l} = 1$ **and** $\sigma^2_{G_l} = 200$



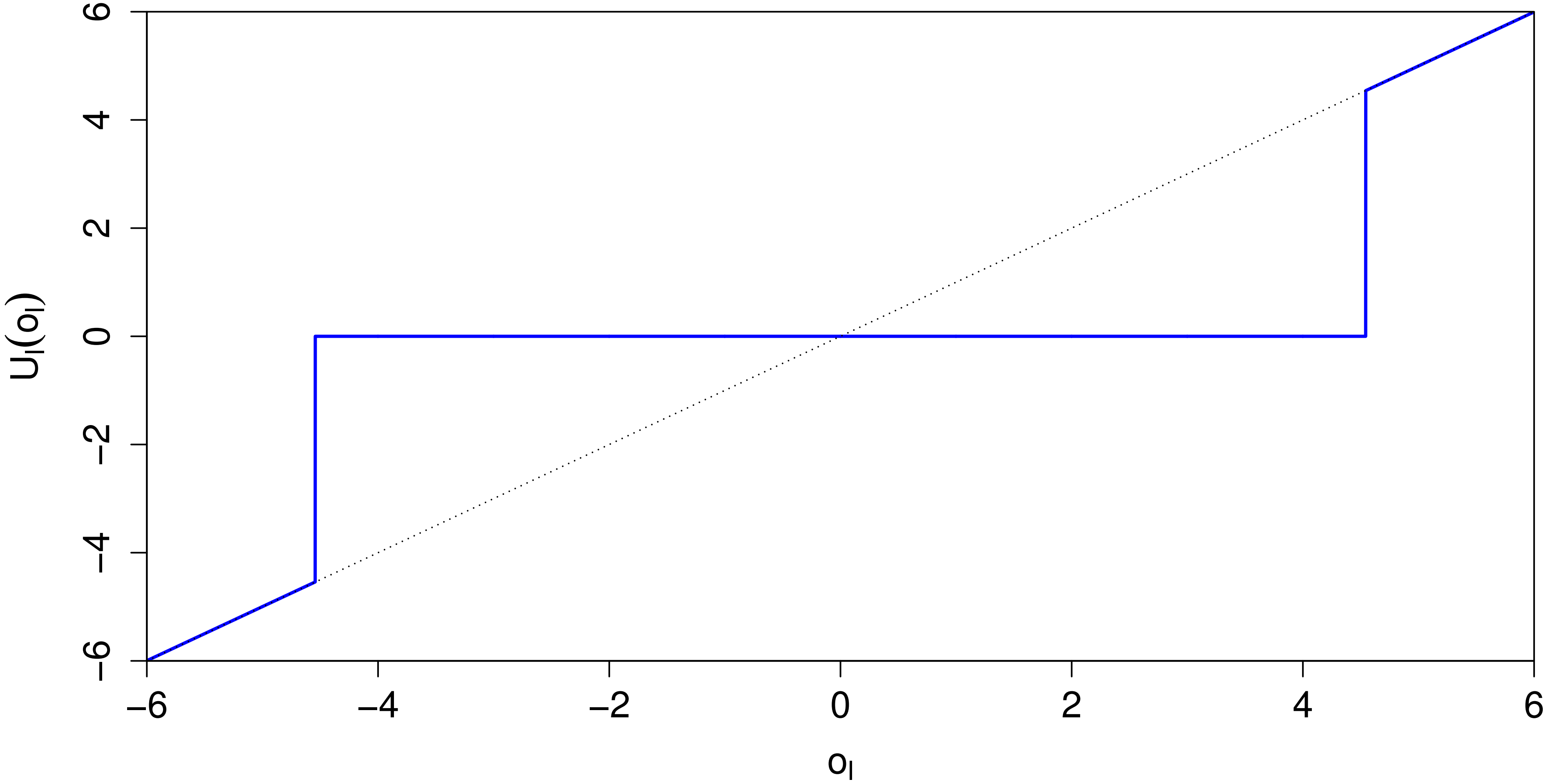
$p_l = 0.95$**,** $\sigma^2_{n_l} = 1$ **and** $\sigma^2_{G_l} = 400$



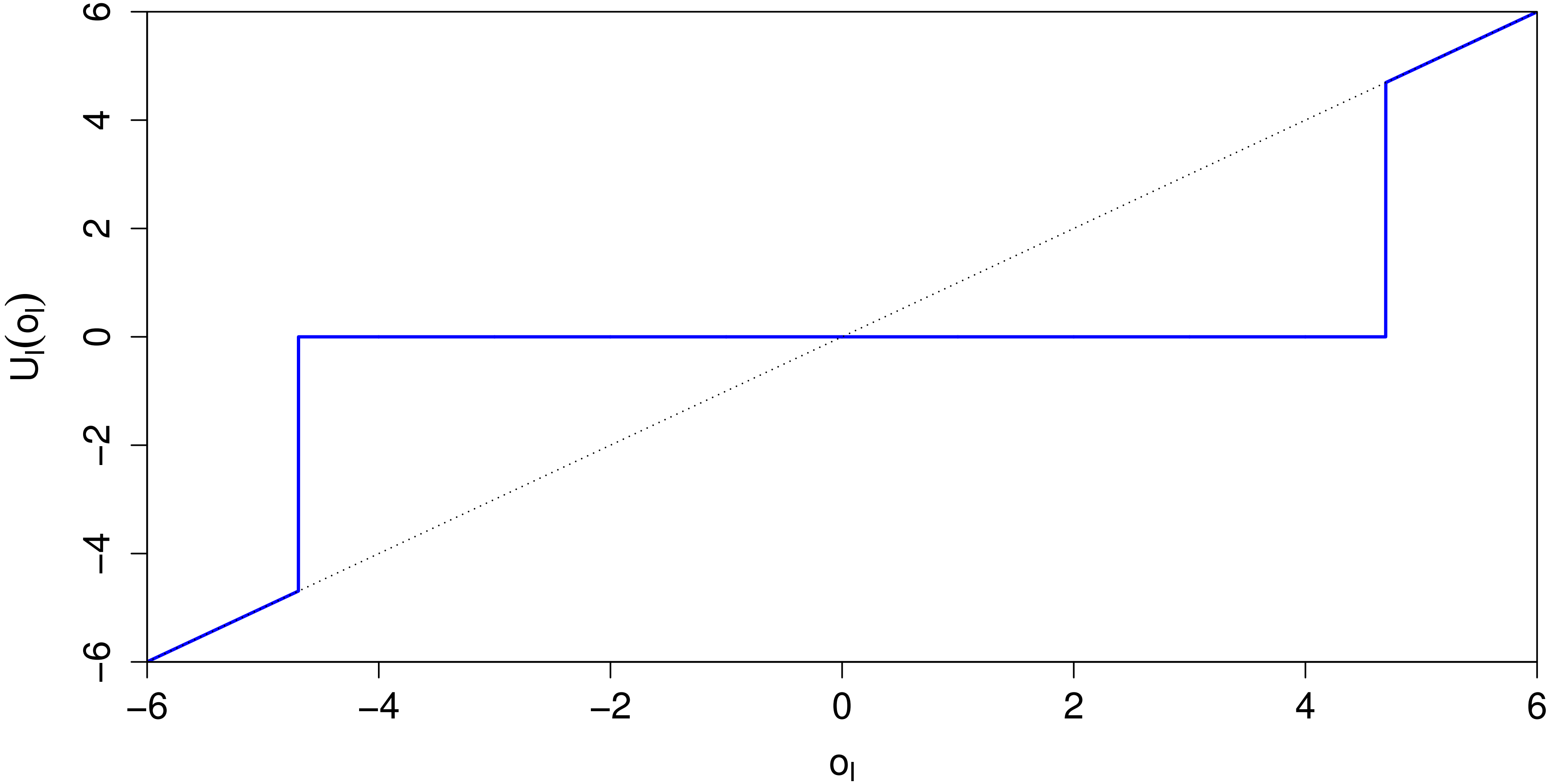
$$p_l = 0.95, \ \sigma^2_{n_l} = 1 \ \textbf{and} \ \sigma^2_{G_l} = 800$$



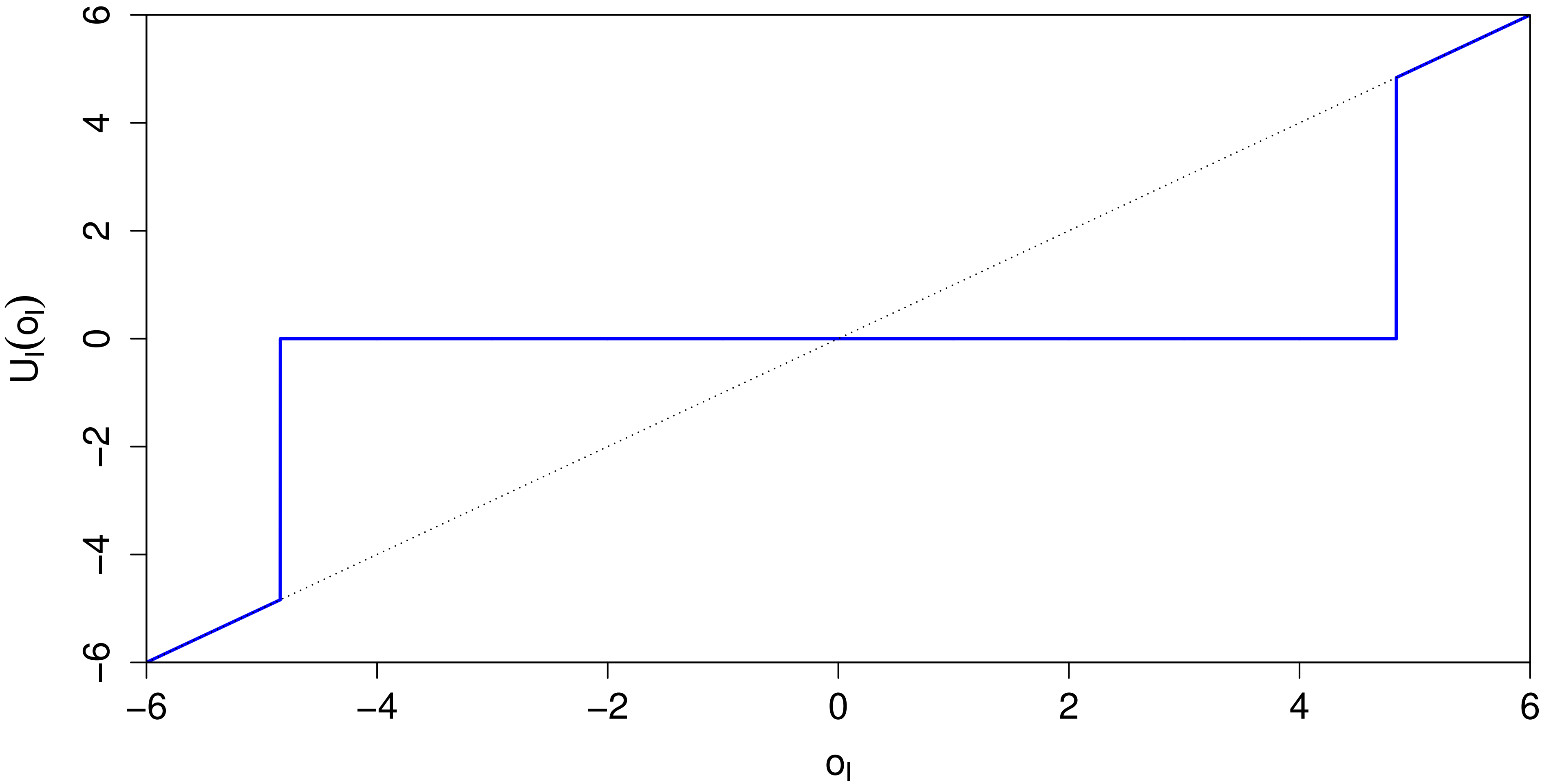
$p_l = 0.95$**,** $\sigma^2_{n_l} = 1$ **and** $\sigma^2_{G_l} = 1600$



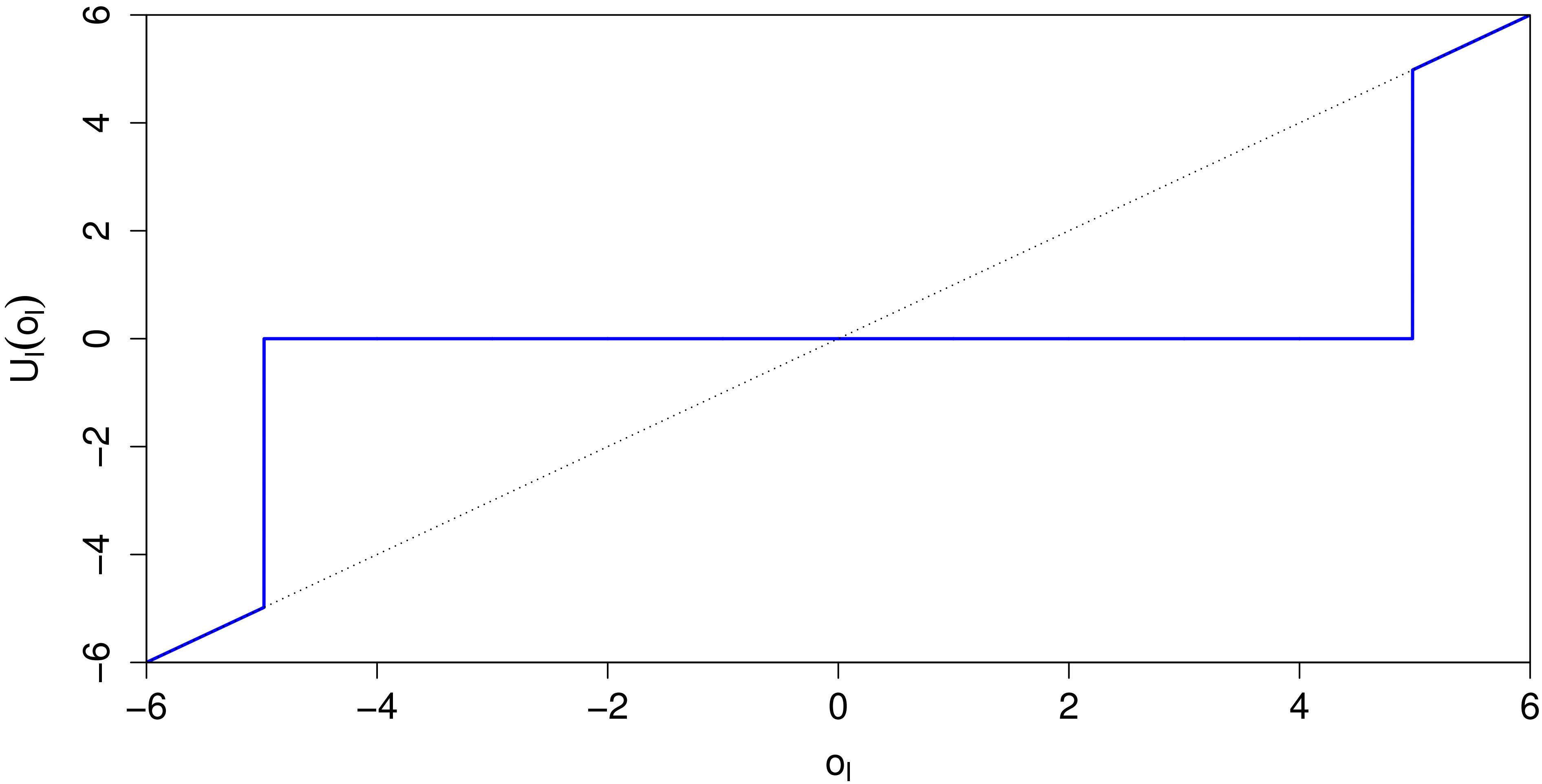
$p_l = 0.95$, $\sigma^2_{n_l} = 1$ **and** $\sigma^2_{G_l} = 3200$



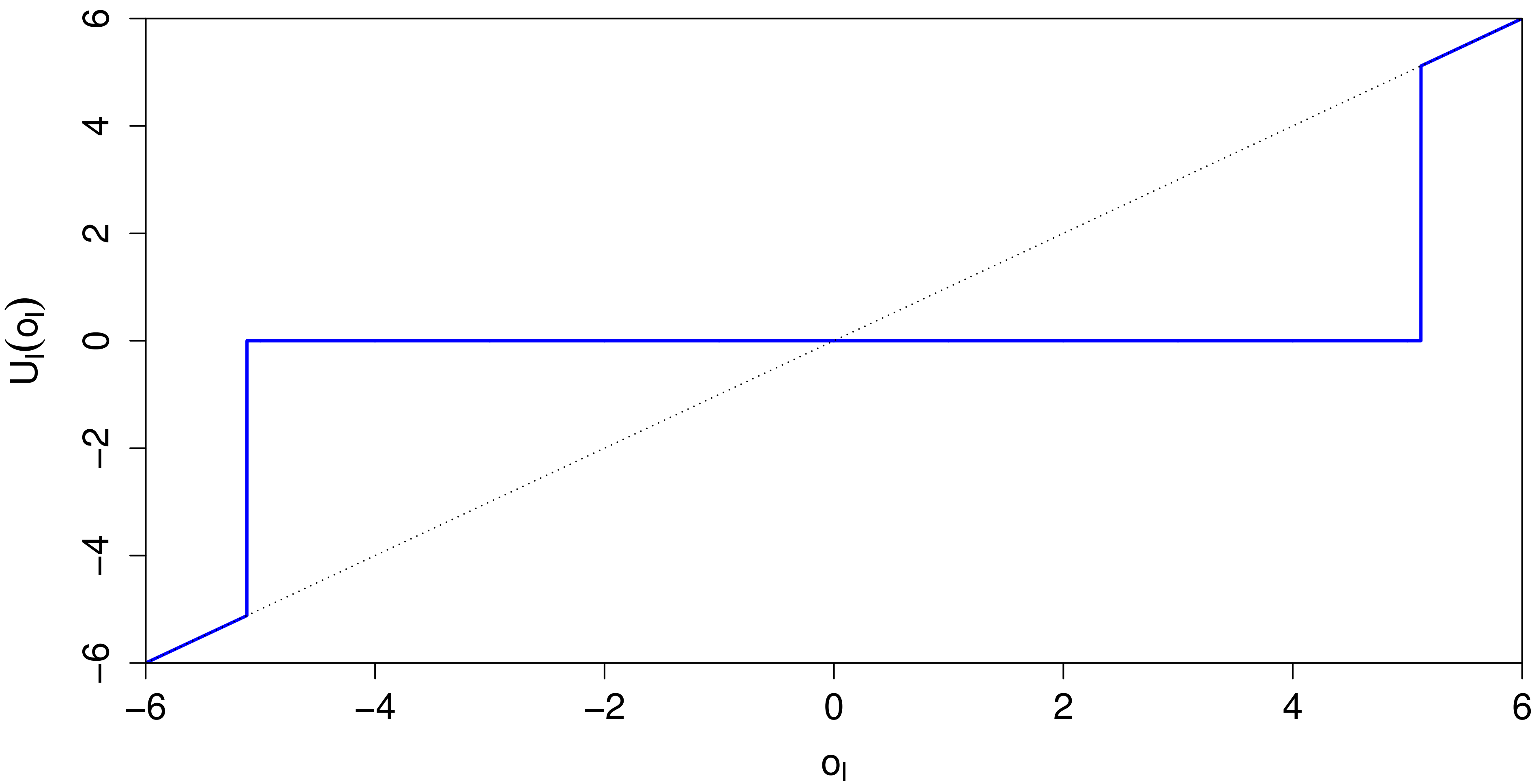
$p_l = 0.95$, $\sigma^2_{n_l} = 1$ **and** $\sigma^2_{G_l} = 6400$



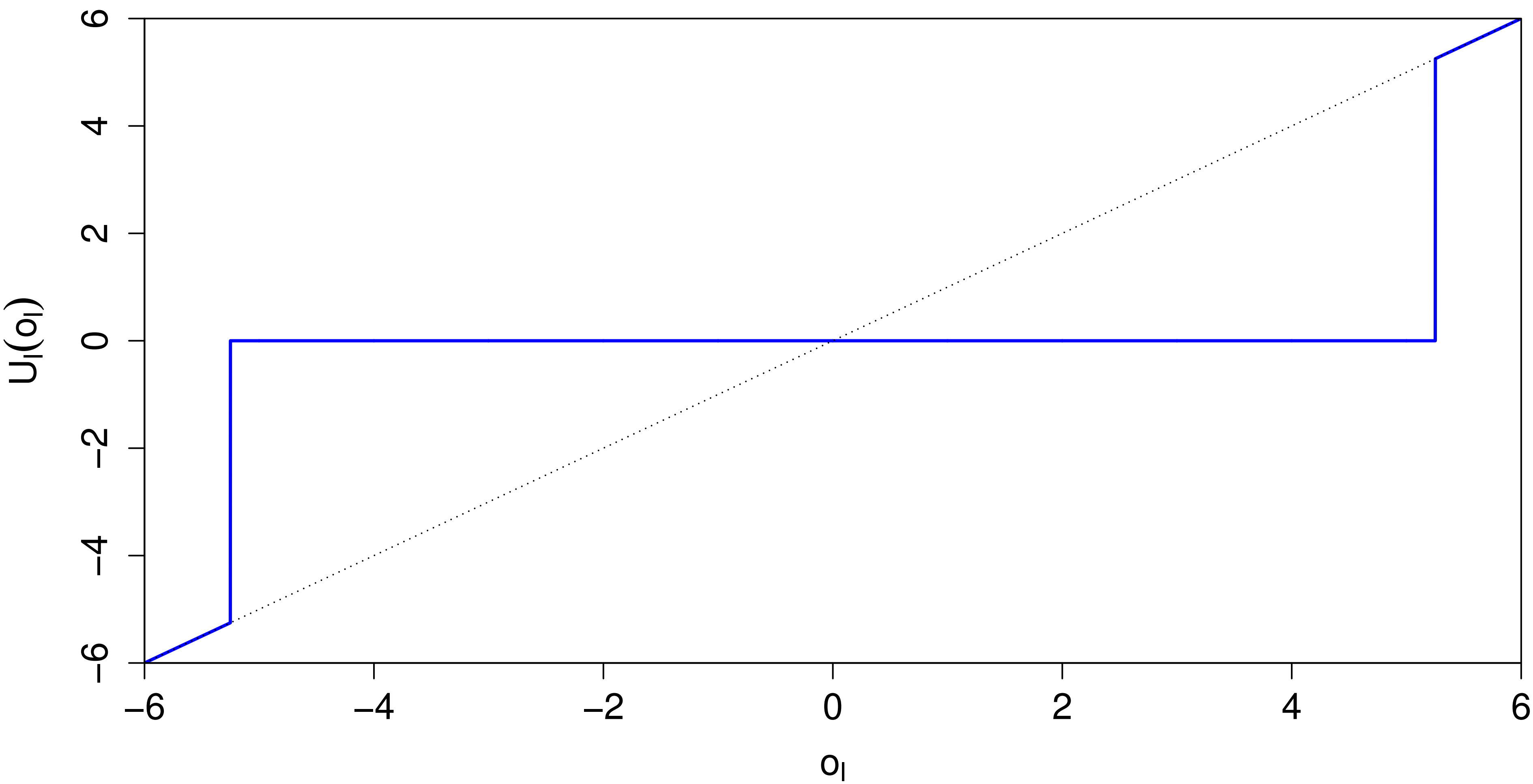
$$p_l = 0.95, \ \sigma^2_{n_l} = 1 \ \textbf{and} \ \sigma^2_{G_l} = 12800$$



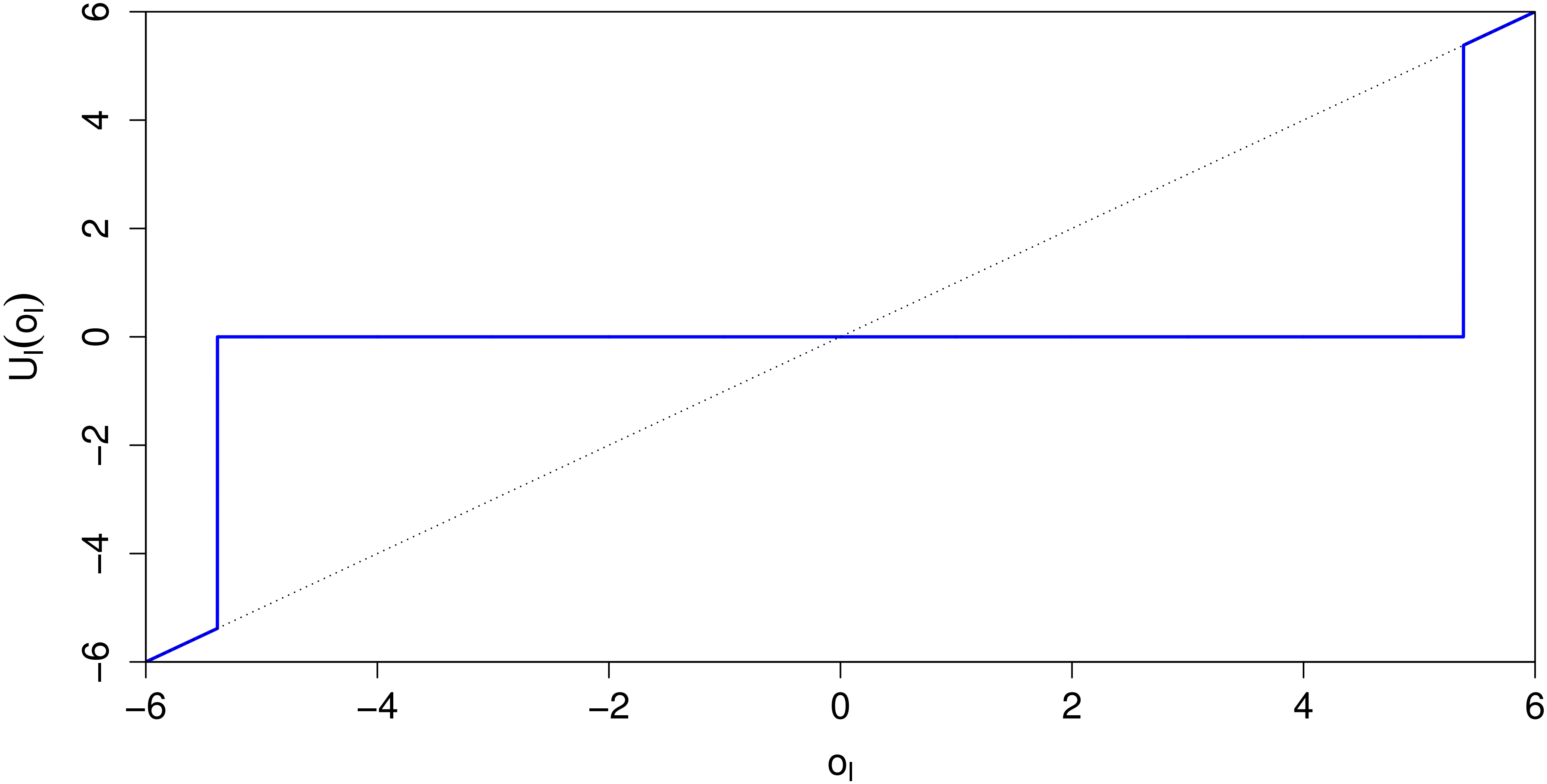
$p_l = 0.95$, $\sigma^2_{n_l} = 1$ **and** $\sigma^2_{G_l} = 25600$



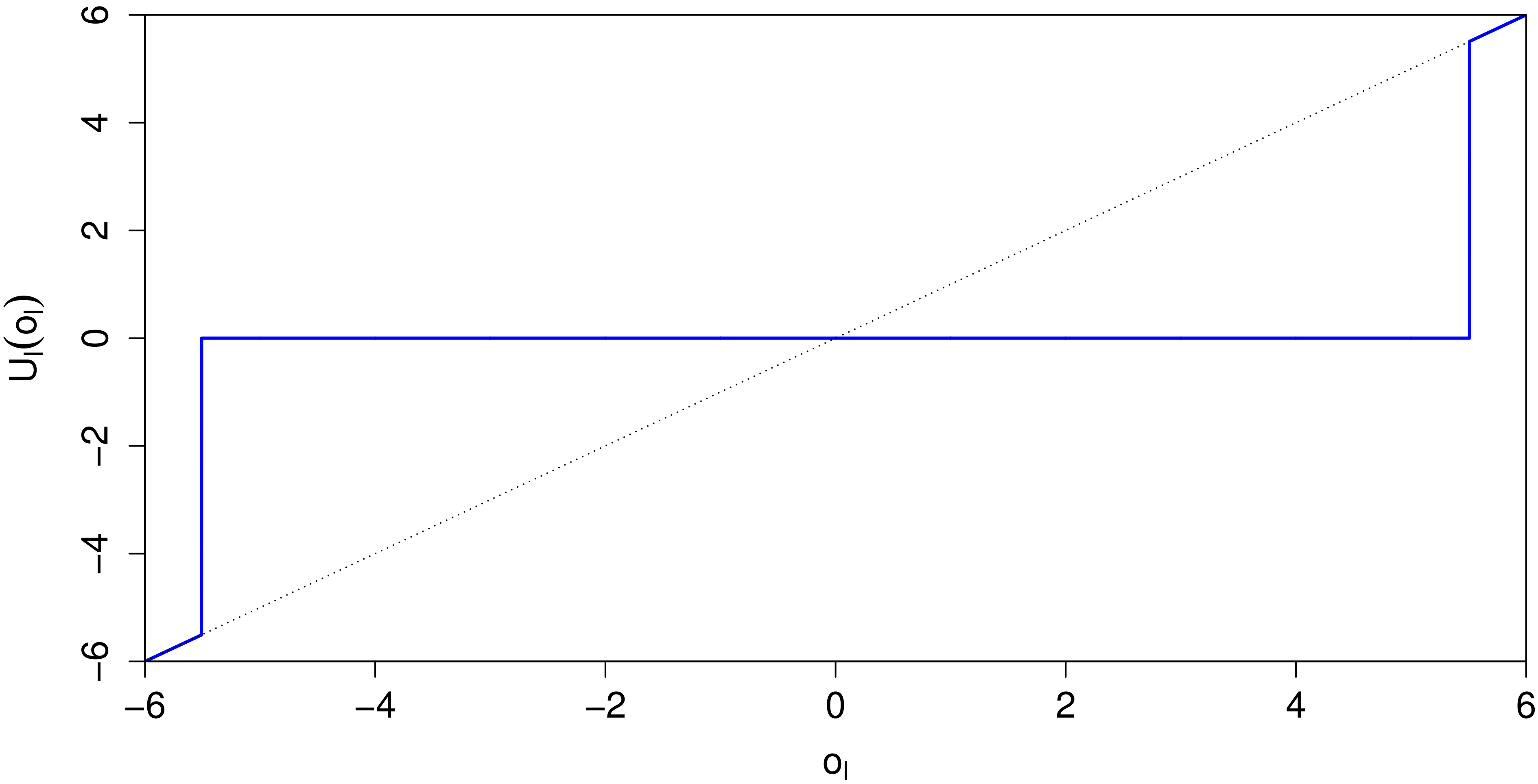
$p_l = 0.95$, $\sigma^2_{n_l} = 1$ **and** $\sigma^2_{G_l} = 51200$



$p_l = 0.95$, $\sigma^2_{n_l} = 1$ **and** $\sigma^2_{G_l} = 102400$



$p_l = 0.95$, $\sigma^2_{n_l} = 1$ **and** $\sigma^2_{G_l} = 204800$



Wavelet-Based Thresholding

- assume model of deterministic signal plus IID Gaussian noise with mean 0 and variance σ_ϵ^2: $\mathbf{X} = \mathbf{D} + \boldsymbol{\epsilon}$
- using a DWT matrix $\mathcal{W}$, form $\mathbf{W} = \mathcal{W}\mathbf{X} = \mathcal{W}\mathbf{D} + \mathcal{W}\boldsymbol{\epsilon} \equiv \mathbf{d} + \mathbf{e}$
- because $\boldsymbol{\epsilon}$ IID Gaussian, so is $\mathbf{e}$ (see Exer. [263])
- Donoho & Johnstone (1994) advocate the following:
  - form partial DWT of level J_0: $\mathbf{W}_1, \ldots, \mathbf{W}_{J_0}$ and $\mathbf{V}_{J_0}$
  - threshold $\mathbf{W}_j$'s but leave $\mathbf{V}_{J_0}$ alone (i.e., administratively, all $N/2^{J_0}$ scaling coefficients assumed to be part of $\mathbf{d}$)
  - use universal threshold $\delta^{(u)} = \sqrt{[2\sigma_\epsilon^2 \log(N)]}$
  - use thresholding rule to form $\mathbf{W}_j^{(t)}$ (hard, etc.)
  - estimate $\mathbf{D}$ by inverse transforming $\mathbf{W}_1^{(t)}, \ldots, \mathbf{W}_{J_0}^{(t)}$ and $\mathbf{V}_{J_0}$

MAD Scale Estimator: I

- procedure assumes σ_ϵ is know, which is not usually the case
- if unknown, use median absolute deviation (MAD) scale estimator to estimate σ_ϵ using $\mathbf{W}_1$

$$\hat{\sigma}_{(\mathrm{mad})} \equiv \frac{\mathrm{median}\,\{|W_{1,0}|, |W_{1,1}|, \ldots, |W_{1,\frac{N}{2}-1}|\}}{0.6745}$$

  - heuristic: bulk of $W_{1,t}$'s should be due to noise
  - '0.6745' yields estimator such that $E\{\hat{\sigma}_{(\mathrm{mad})}\} = \sigma_\epsilon$ when $W_{1,t}$'s are IID Gaussian with mean 0 and variance σ_ϵ^2
  - designed to be robust against large $W_{1,t}$'s due to signal

MAD Scale Estimator: II

- example: suppose $\mathbf{W}_1$ has 7 small 'noise' coefficients & 2 large 'signal' coefficients (say, a & b, with $2 \ll |a| < |b|$):

$$\mathbf{W}_1 = [1.23, -1.72, -0.80, -0.01, a, 0.30, 0.67, b, -1.33]^T$$

- ordering these by their magnitudes yields

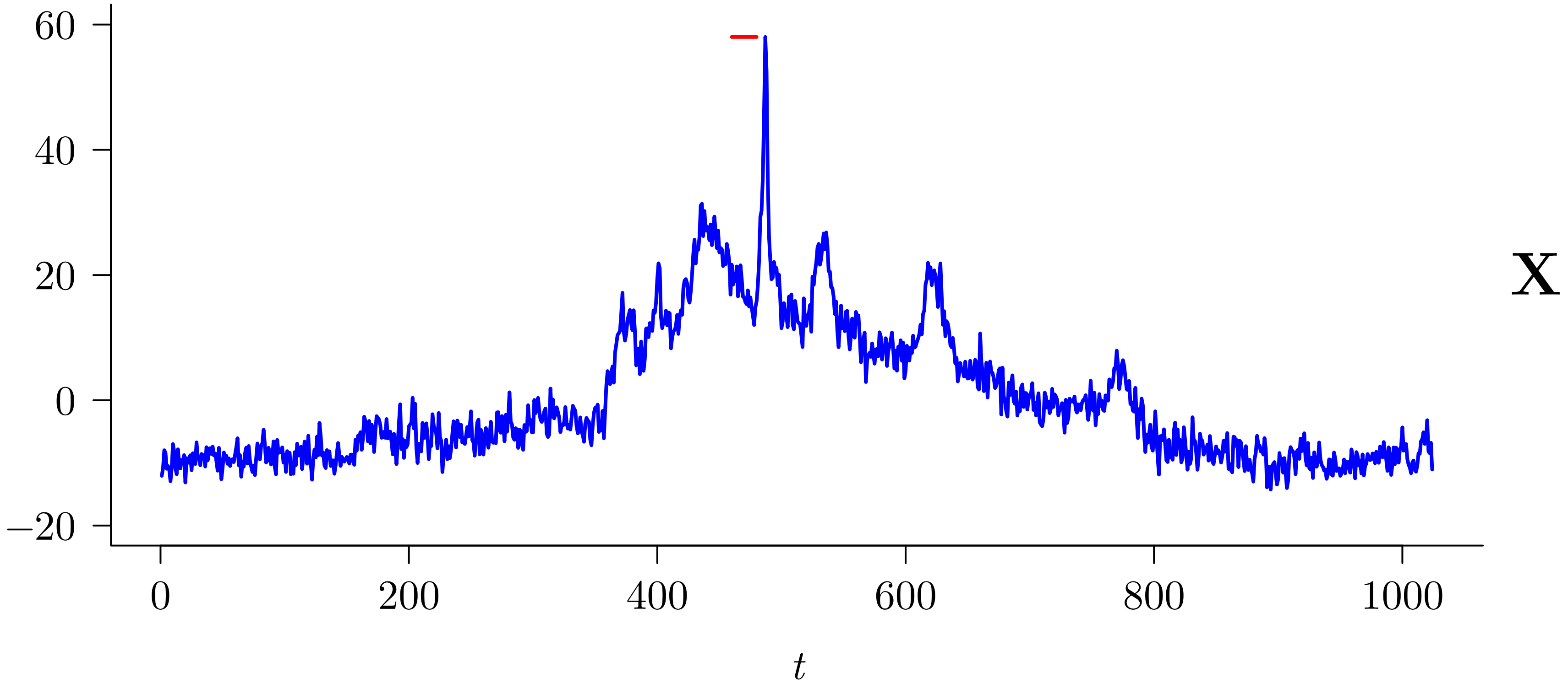
$$0.01, 0.30, 0.67, 0.80, 1.23, 1.33, 1.72, |a|, |b|$$

- median of these absolute deviations is 1.23, so

$$\hat{\sigma}_{(\mathrm{mad})} = 1.23/0.6745 \doteq 1.82$$

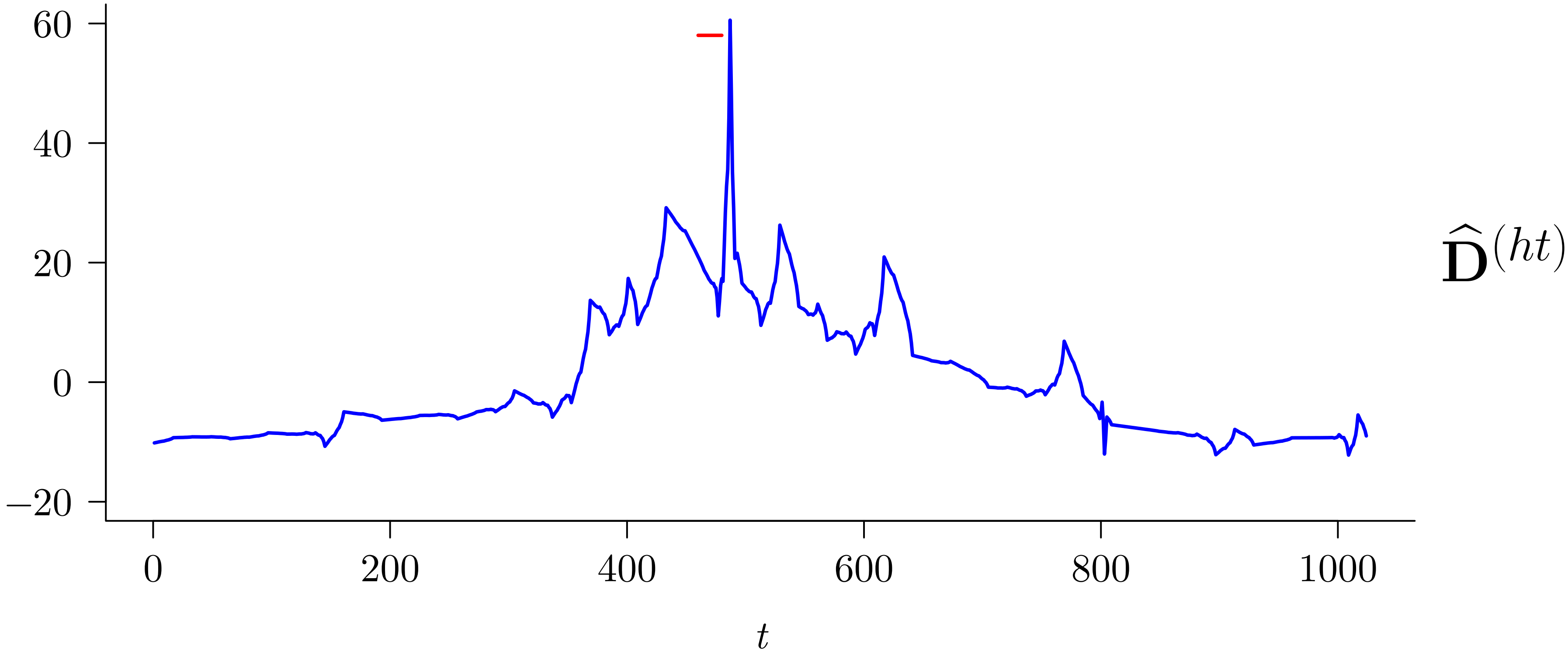
- $\hat{\sigma}_{(\mathrm{mad})}$ not influenced adversely by a and b; i.e., scale estimate depends largely on the many small coefficients due to noise

Examples of DWT-Based Thresholding: I



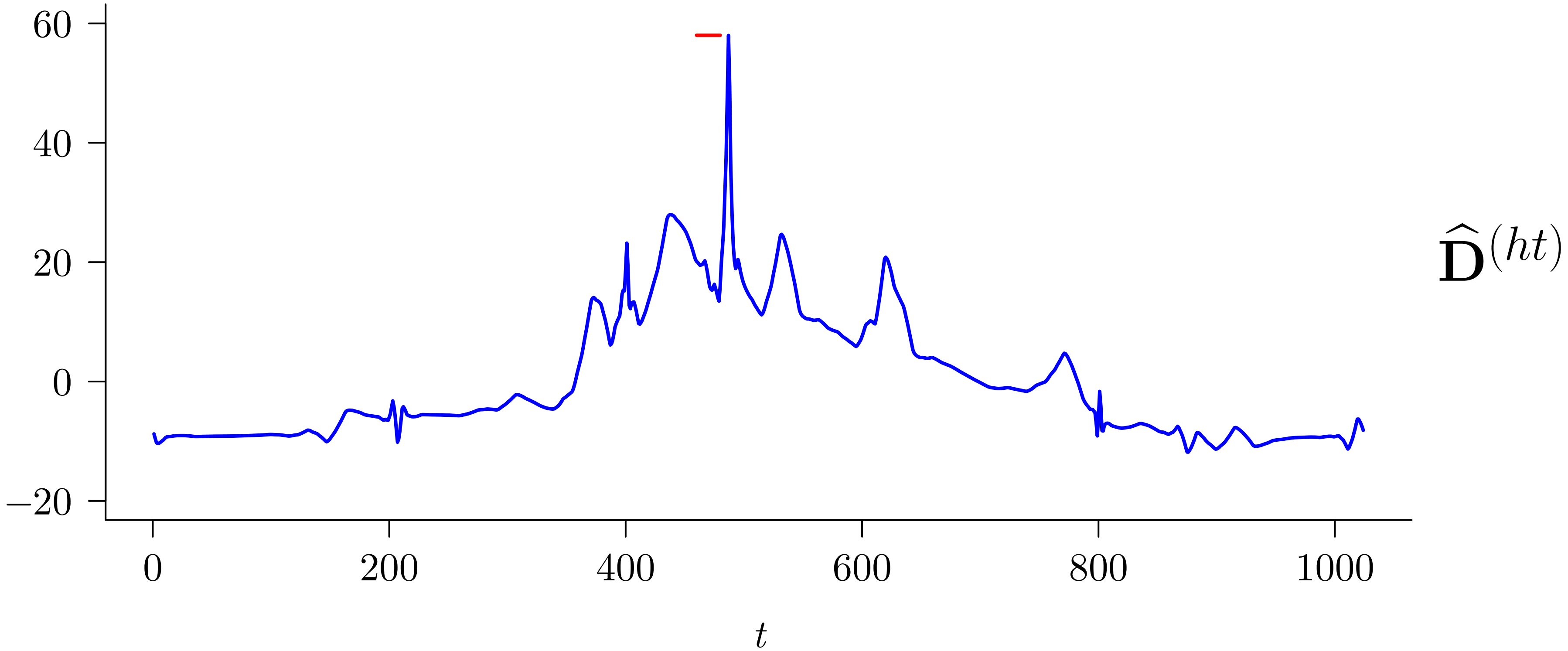
- NMR spectrum

Examples of DWT-Based Thresholding: II



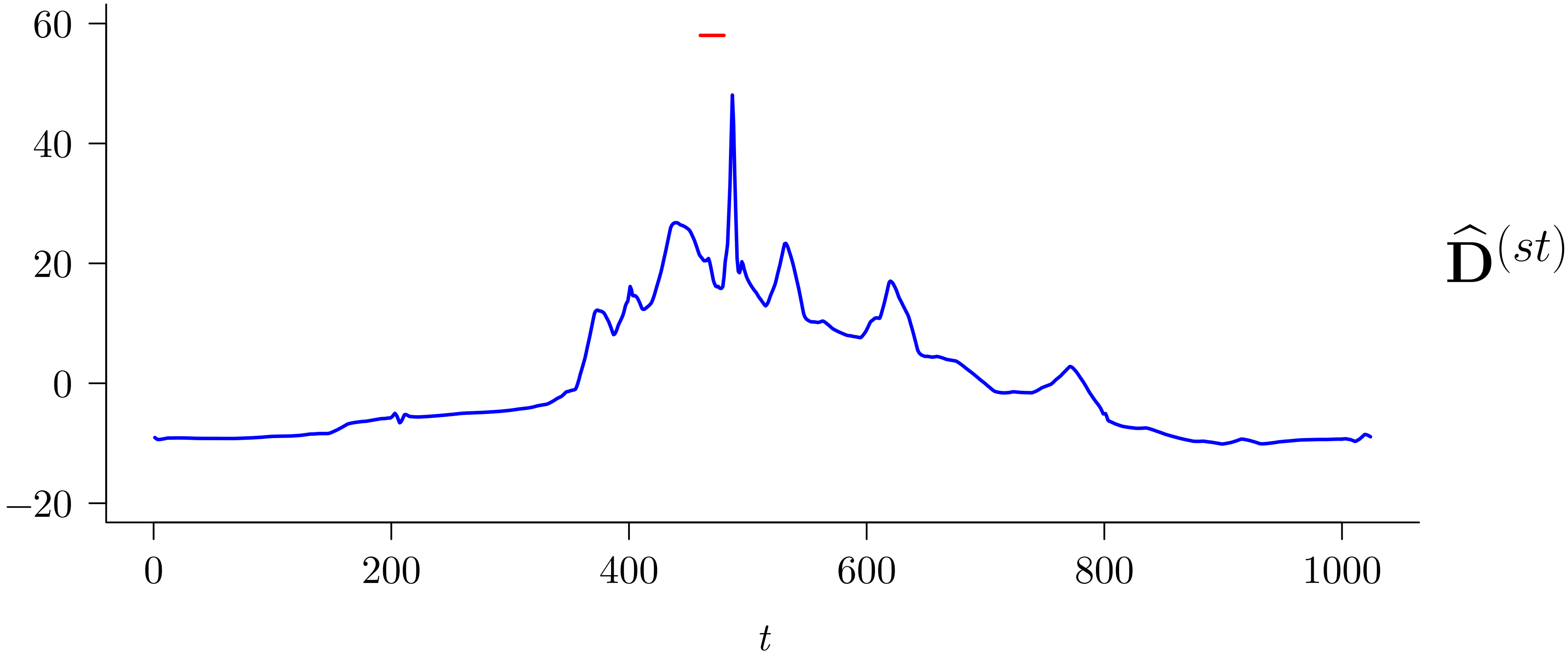
- signal estimate using $J_0 = 6$ partial D(4) DWT with hard thresholding and universal threshold level estimated by $\hat{\delta}^{(u)} = \sqrt{[2\hat{\sigma}^2_{(\text{mad})} \log(N)]} \doteq 6.49$

Examples of DWT-Based Thresholding: III



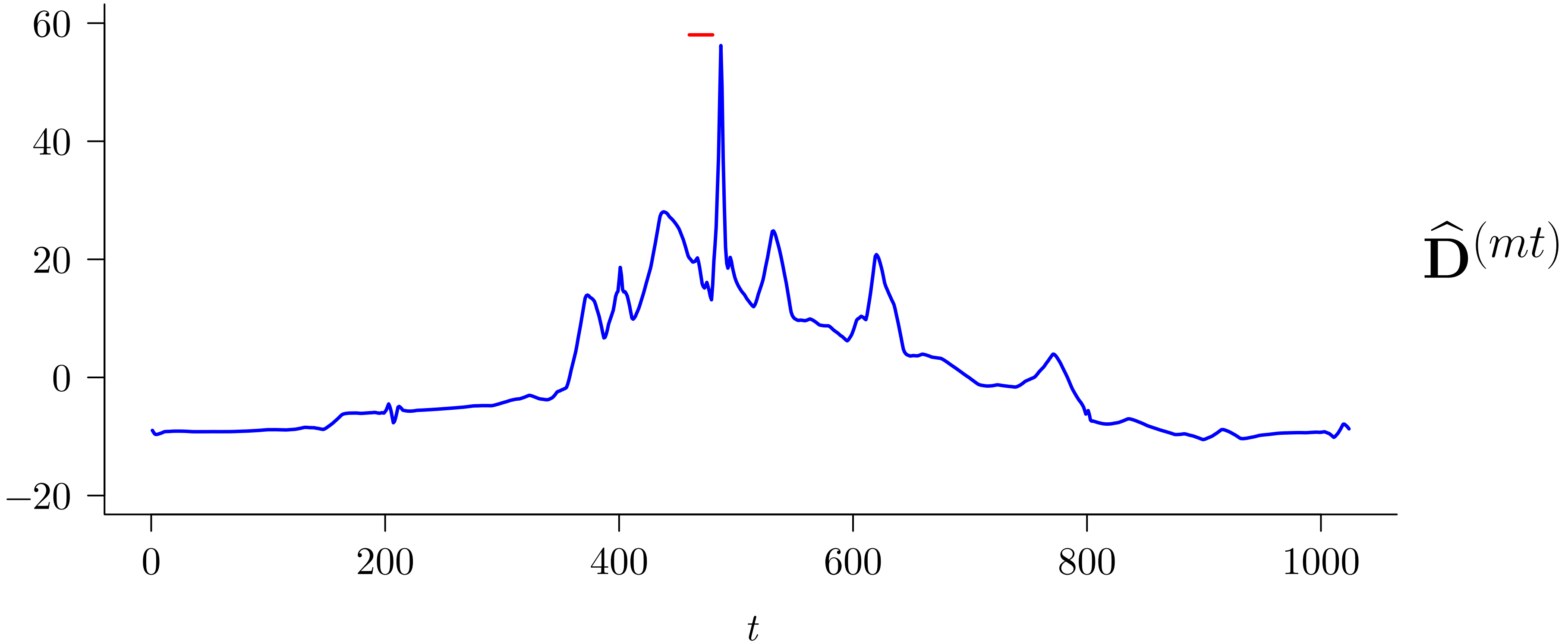
- same as before, but now using LA(8) DWT with $\hat{\delta}^{(u)} \doteq 6.13$

Examples of DWT-Based Thresholding: IV



- signal estimate using $J_0 = 6$ partial LA(8) DWT, but now with soft thresholding

Examples of DWT-Based Thresholding: V



- signal estimate using $J_0 = 6$ partial LA(8) DWT, but now with mid thresholding

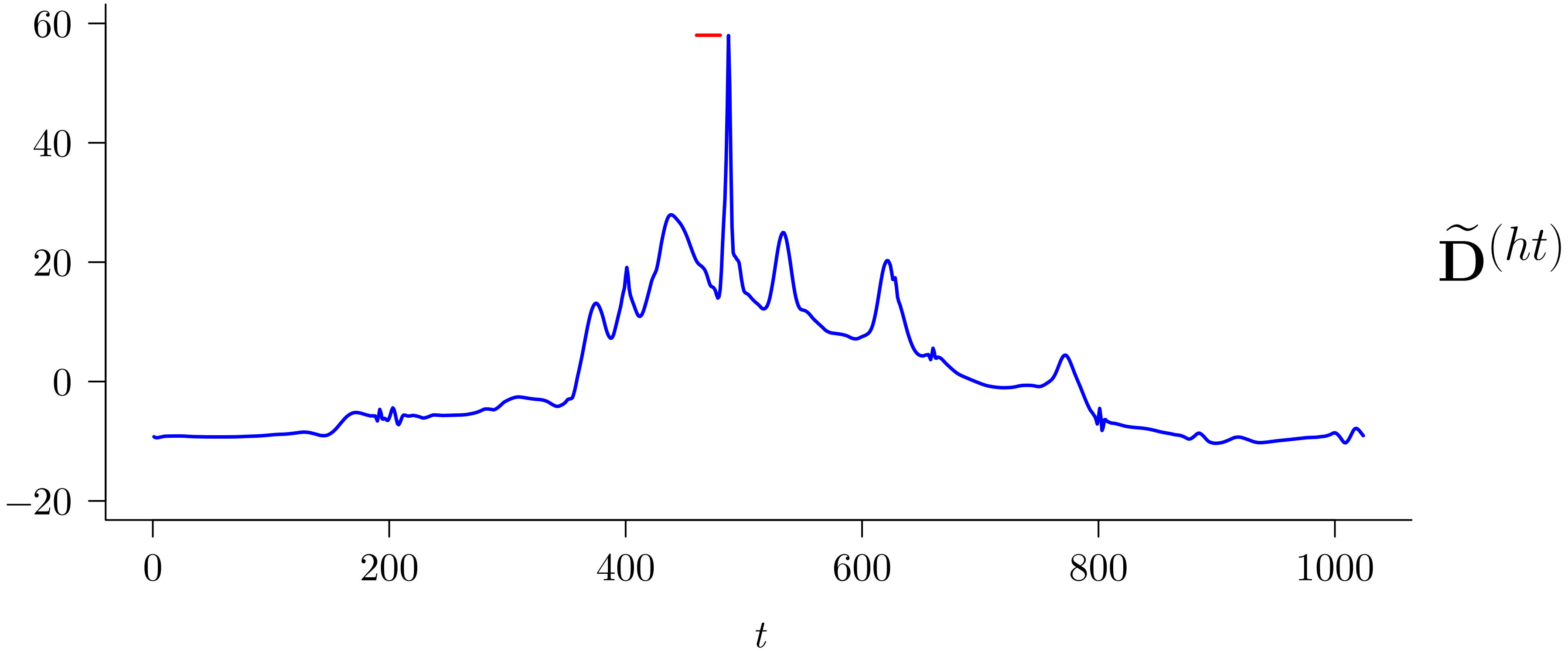
MODWT-Based Thresholding

- can base thresholding procedure on MODWT rather than DWT, yielding signal estimators $\widetilde{\mathbf{D}}^{(ht)}$, $\widetilde{\mathbf{D}}^{(st)}$ and $\widetilde{\mathbf{D}}^{(mt)}$
- because MODWT filters are normalized differently, universal threshold must be adjusted for each level:

$$\tilde{\delta}_j^{(u)} \equiv \sqrt{[\tilde{\sigma}^2_{(\text{mad})} \log{(N)}/2^{j-1}]},$$

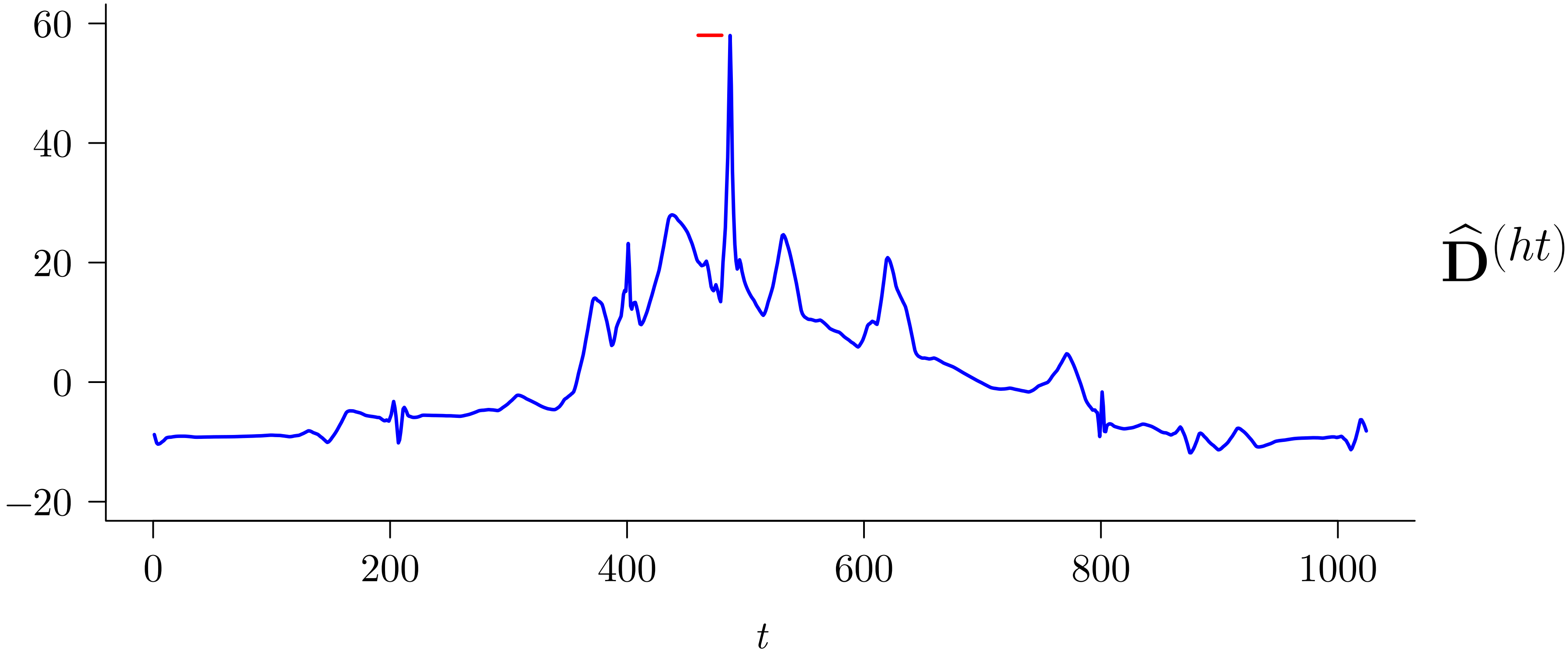
  where now MAD scale estimator is based on unit scale MODWT wavelet coefficients
- results are almost the same as what 'cycle spinning' would yield
  - would be the same if DWT-based MAD estimates $\hat{\sigma}^2_{(\text{mad})}$ were identical for odd/even downsampling and if MODWT-based estimate $\tilde{\sigma}^2_{(\text{mad})}$ were such that $2\tilde{\sigma}^2_{(\text{mad})} = \hat{\sigma}^2_{(\text{mad})}$

Examples of MODWT-Based Thresholding: I



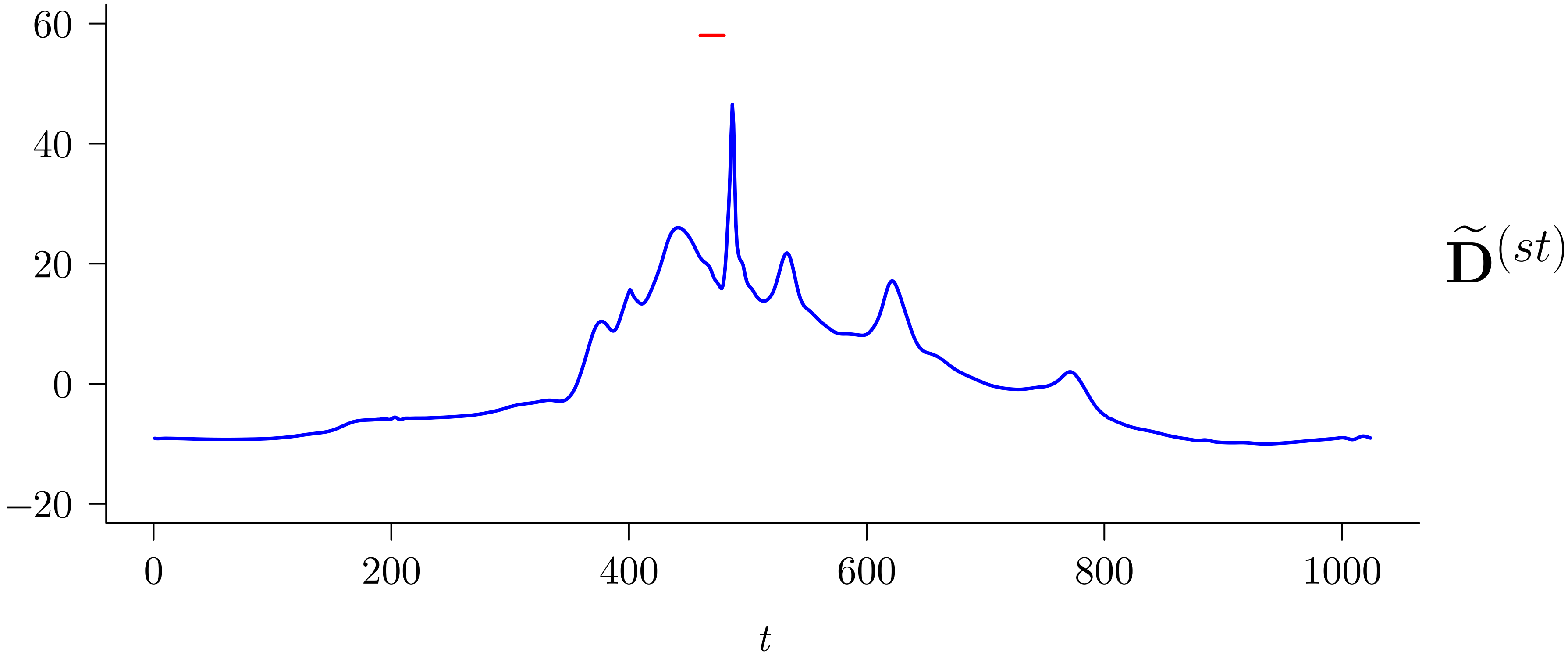
- signal estimate using $J_0 = 6$ LA(8) MODWT with hard thresholding

Examples of DWT-Based Thresholding: III



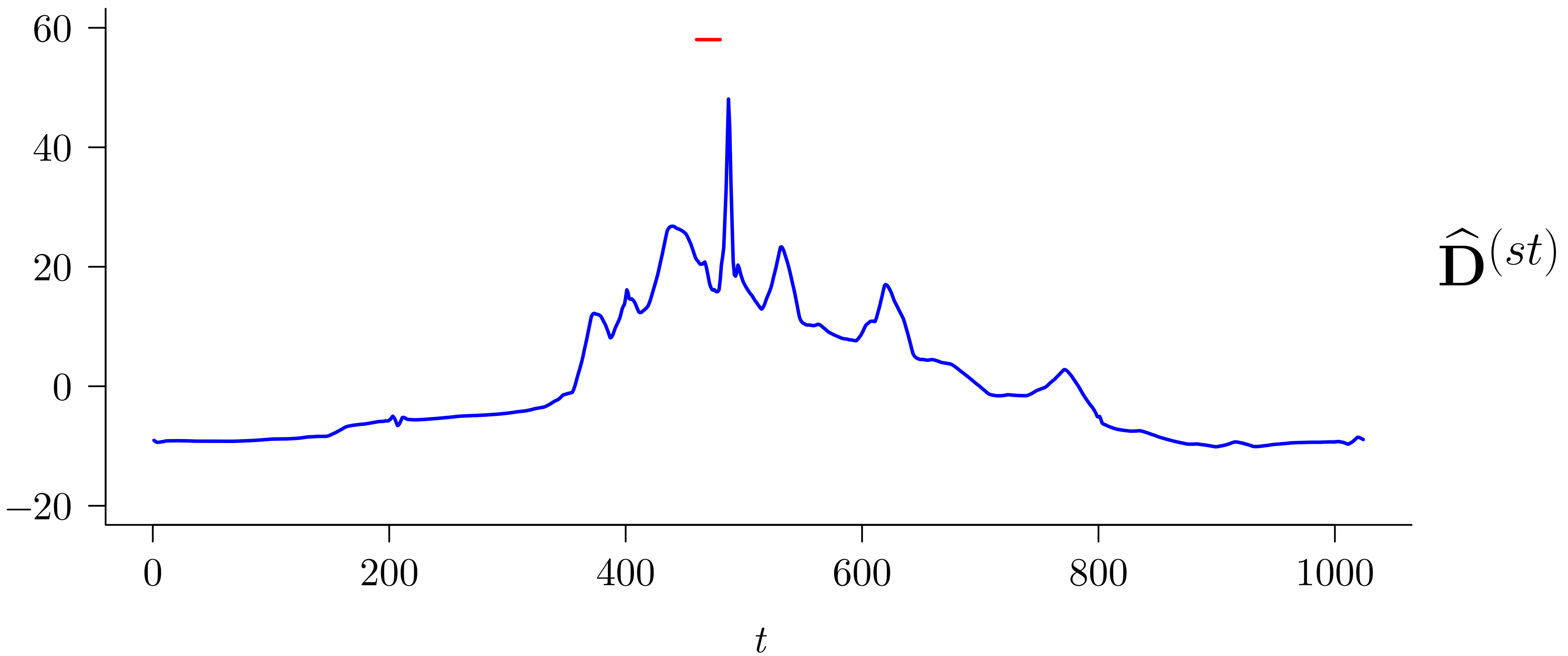
- same as before, but now using LA(8) DWT with $\hat{\delta}^{(u)} \doteq 6.13$

Examples of MODWT-Based Thresholding: II



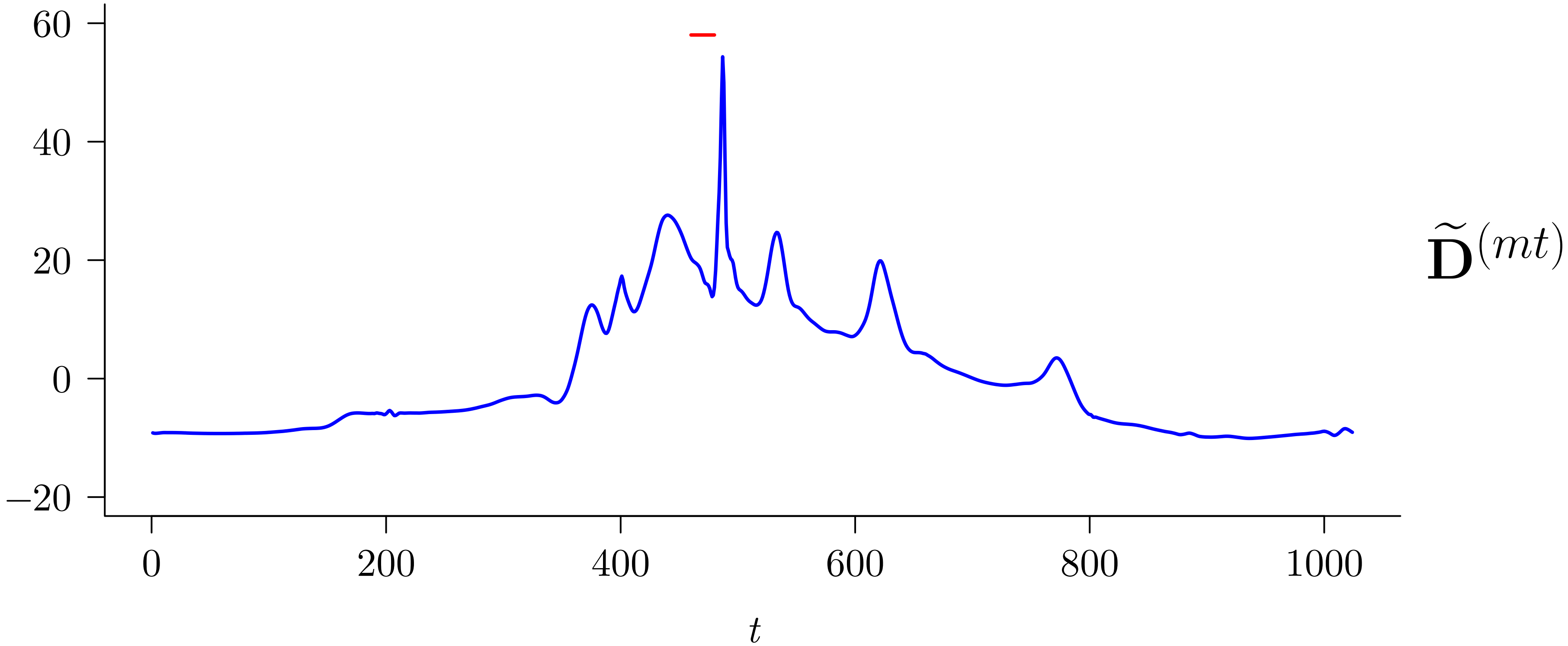
- same as before, but now with soft thresholding

Examples of DWT-Based Thresholding: IV



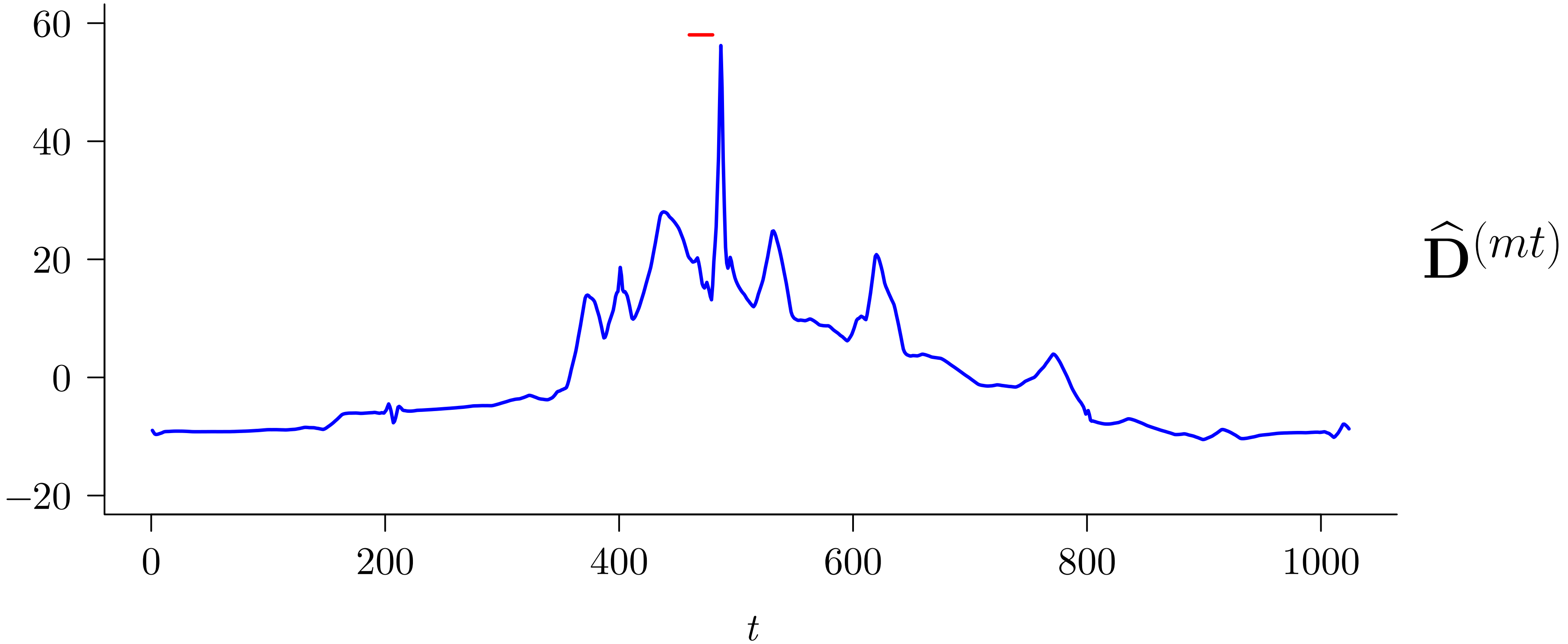
- signal estimate using $J_0 = 6$ partial LA(8) DWT, but now with soft thresholding

Examples of MODWT-Based Thresholding: III



- same as before, but now with mid thresholding

Examples of DWT-Based Thresholding: V



- signal estimate using $J_0 = 6$ partial LA(8) DWT, but now with mid thresholding

VisuShrink: I

- Donoho & Johnstone (1994) recipe with soft thresholding is known as 'VisuShrink' (but really thresholding, not shrinkage)
- one theoretical justification for VisuShrink
  - consider the risk for all possible signals $\mathbf{D}$ using VisuShrink:
  $$R(\widehat{\mathbf{D}}^{(st)}, \mathbf{D}) \equiv E\{\|\widehat{\mathbf{D}}^{(st)} - \mathbf{D}\|^2\}$$
  - consider 'ideal' risk $R(\widehat{\mathbf{D}}^{(i)}, \mathbf{D})$ formed with the help of an 'oracle' that tells us which $W_{j,t}$'s are dominated by noise
  - Donoho & Johnstone (1994), Theorem 1:
  $$R(\widehat{\mathbf{D}}^{(st)}, \mathbf{D}) \leq [2\log(N) + 1][\sigma_\epsilon^2 + R(\widehat{\mathbf{D}}^{(i)}, \mathbf{D})]$$
  - two risks differ by only a logarithmic factor
  - risks for other estimators do poorer when compared to the 'ideal' risk

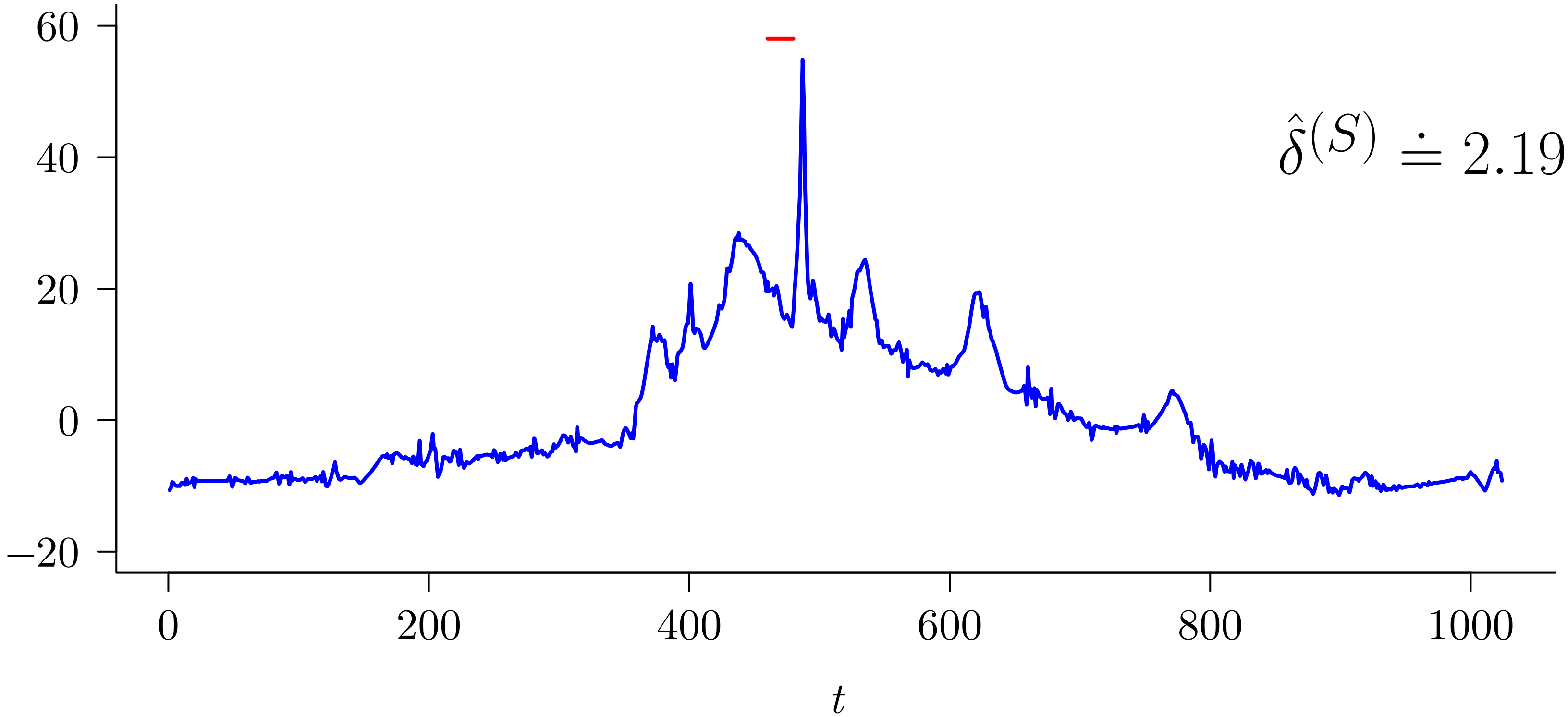
VisuShrink: II

- rather than using the universal threshold, can also determine δ for VisuShrink by finding value $\hat{\delta}^{(S)}$ that minimizes SURE, i.e.,

$$\sum_{j=1}^{J_0} \sum_{t=0}^{N_j-1} (2\hat{\sigma}^2_{(\text{mad})} - W^2_{j,t} + \delta^2) 1_{[\delta^2,\infty)}(W^2_{j,t}),$$

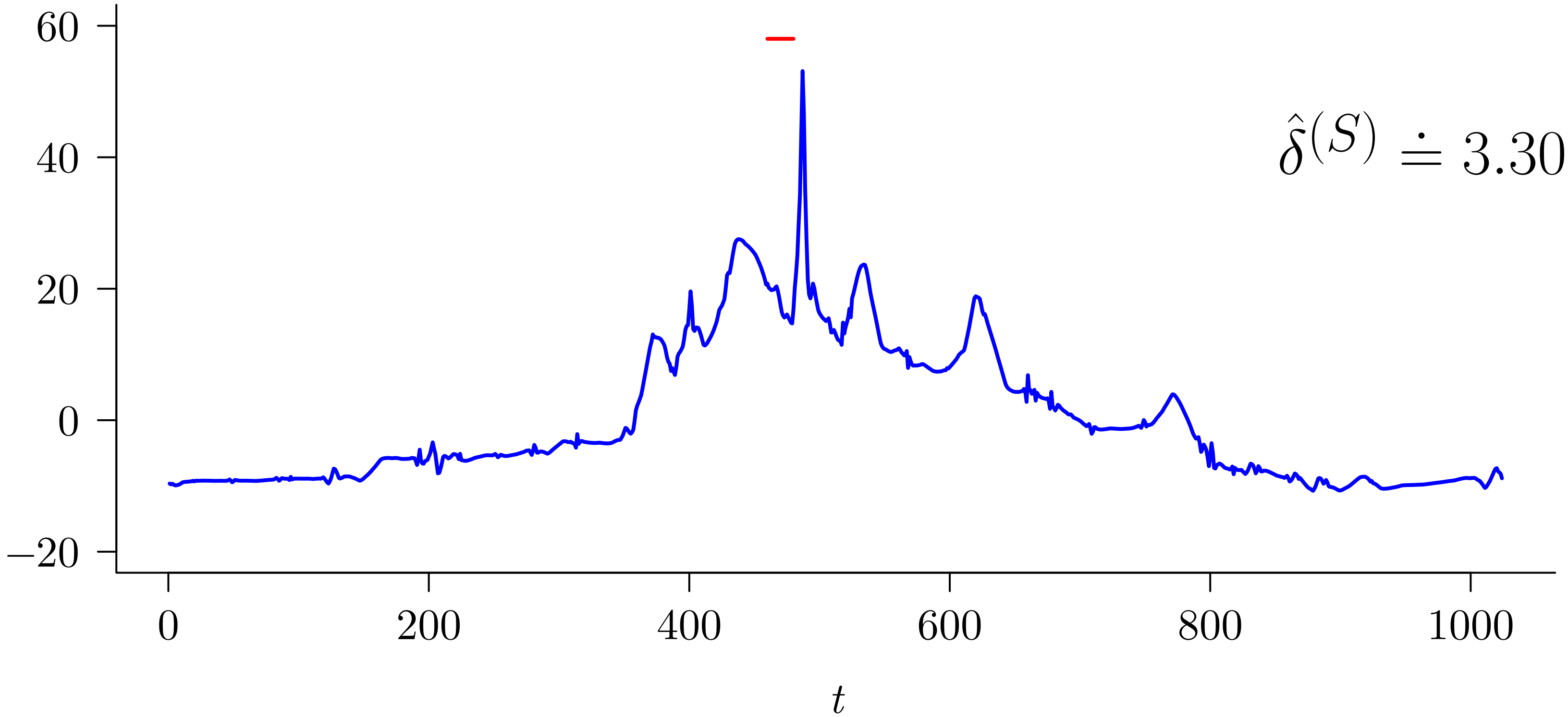
as a function of δ, with σ^2_ϵ estimated via MAD

Examples of DWT-Based Thresholding: III



- VisuShrink estimate based upon level $J_0 = 6$ partial LA(8) DWT and SURE with MAD estimate based upon $\mathbf{W}_1$

Examples of DWT-Based Thresholding: IV



- same as before, but now with MAD estimate based upon $\mathbf{W}_1$, $\mathbf{W}_2, \ldots, \mathbf{W}_6$ (the common variance in SURE is assumed common to all wavelet coefficients) – signal estimate less noisy

Wavelet-Based Shrinkage: I

- assume model of stochastic signal plus Gaussian IID noise: $\mathbf{X} = \mathbf{C} + \boldsymbol{\epsilon}$ so that $\mathbf{W} = \mathcal{W}\mathbf{X} = \mathcal{W}\mathbf{C} + \mathcal{W}\boldsymbol{\epsilon} \equiv \mathbf{R} + \mathbf{e}$
- component-wise, have $W_{j,t} = R_{j,t} + e_{j,t}$, with $R_{j,t}$ & $e_{j,t}$ being independent RVs, both with zero means
- form partial DWT of level J_0, shrink $\mathbf{W}_j$'s, but leave $\mathbf{V}_{J_0}$ alone (assumption $E\{R_{j,t}\} = 0$ reasonable for $\mathbf{W}_j$, but not for $\mathbf{V}_{J_0}$)
- use conditional mean approach
  - $R_{j,t}$'s are IID with distribution given by $(1 - \mathcal{I}_{j,t})\mathcal{N}(0, \sigma_G^2)$, i.e., a sparse signal model, where

$$\mathbf{P}\left[\mathcal{I}_{j,t} = 1\right] = p \text{ and } \mathbf{P}\left[\mathcal{I}_{j,t} = 0\right] = 1 - p$$

  - $e_{j,t}$ has distribution dictated by $\mathcal{N}(0, \sigma_\epsilon^2)$
  - note: parameters do not vary with j or t

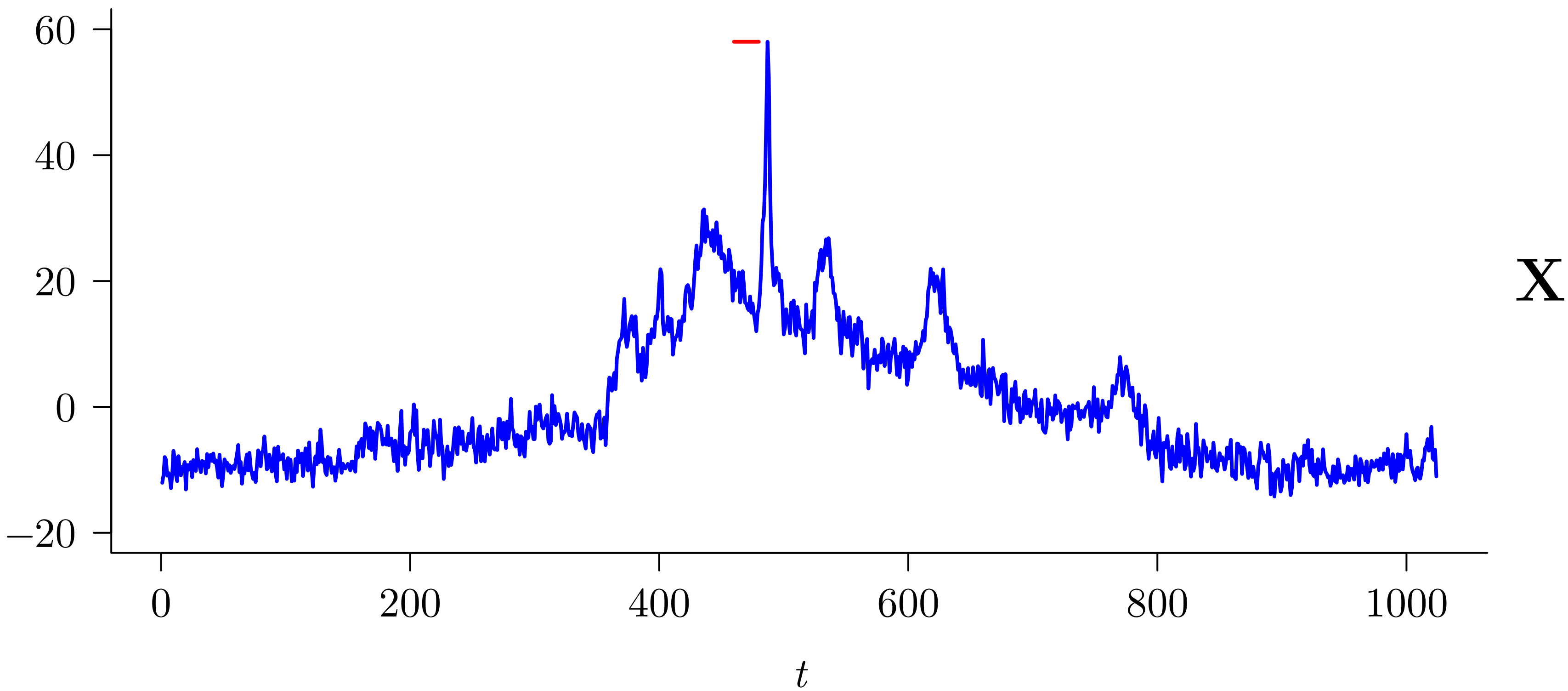
Wavelet-Based Shrinkage: II

- model has three parameters that need to be set, two related to signal (σ_G^2 & p), and one related to noise (σ_ϵ^2)
- can use $\mathbf{W}_1$ to estimate σ_ϵ^2 via $\hat{\sigma}_\epsilon^2 = \hat{\sigma}_{(\mathrm{mad})}^2$
- wavelet coefficients in $\mathbf{W}_1, \ldots, \mathbf{W}_{J_0}$ have a common variance σ_W^2, which can be estimated by sample mean $\hat{\sigma}_W^2$ of all $W_{j,t}^2$'s
- can use relationship

$$\sigma_G^2 = \frac{\sigma_W^2 - \sigma_\epsilon^2}{1 - p}$$

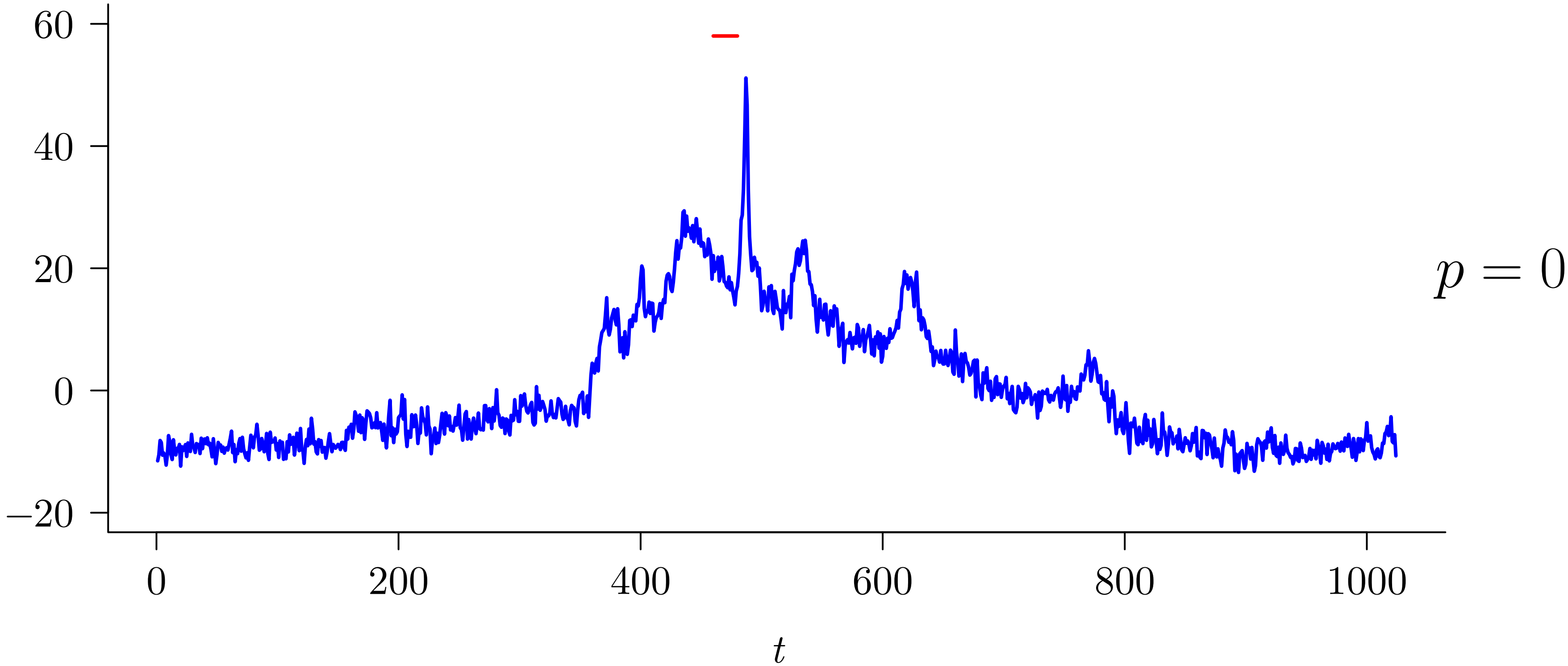
to create estimator $\hat{\sigma}_G^2$ once p is chosen (usually subjectively, but keeping in mind that p is proportion of noise-dominated coefficients – might be able to set based on rough estimate of proportion of 'small' coefficients)

Examples of Wavelet-Based Shrinkage: I



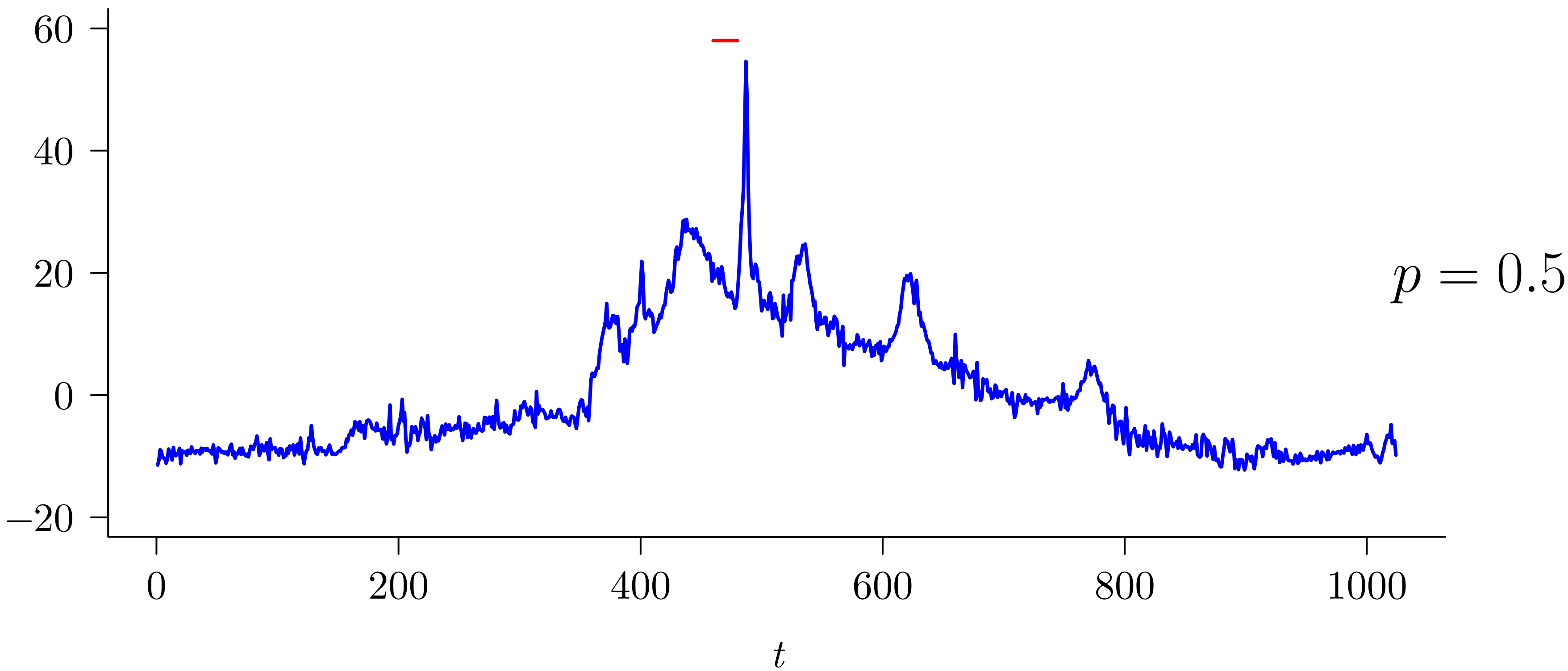
- NMR spectrum

Examples of Wavelet-Based Shrinkage: II



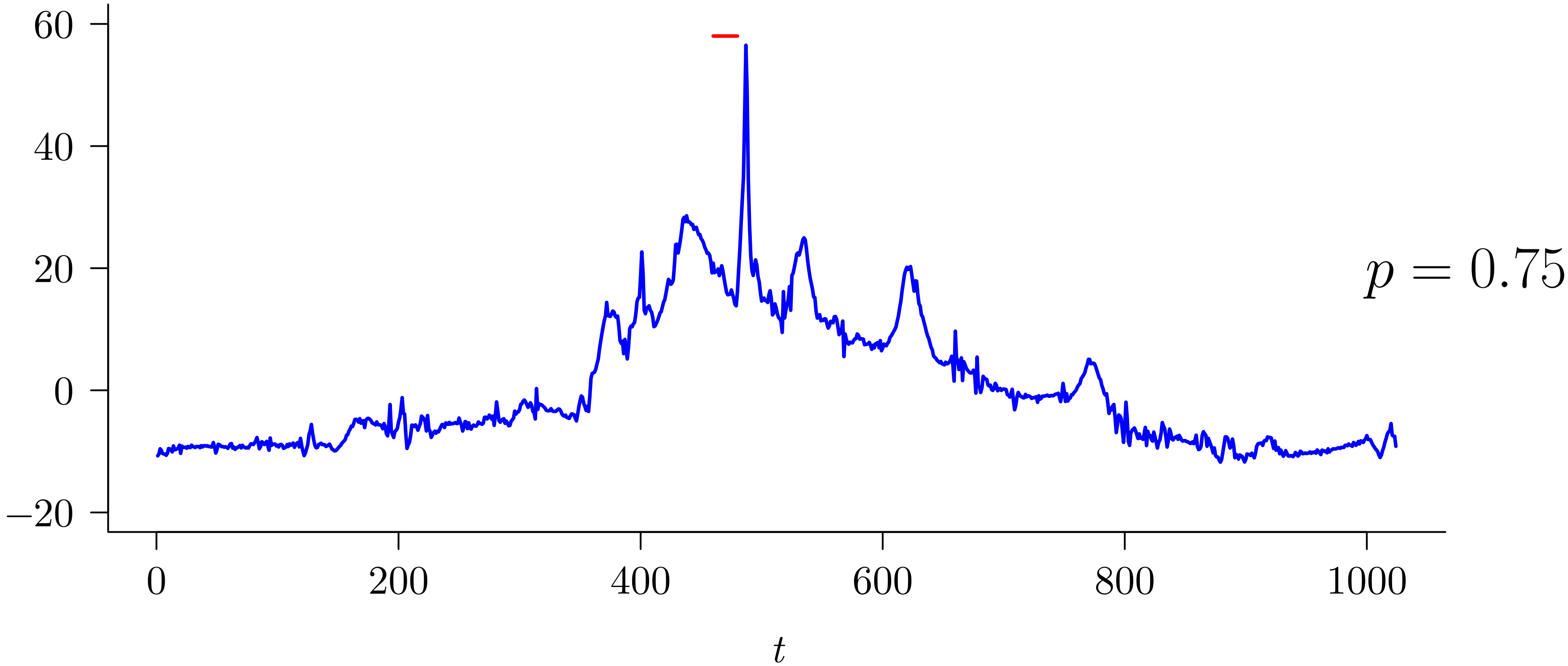
- shrinkage signal estimates of NMR spectrum based upon level $J_0 = 6$ partial LA(8) DWT and conditional mean with $p = 0$ (with this choice of p, estimator collapses to minimum mean square estimator of ovehead XI–37)

Examples of Wavelet-Based Shrinkage: III



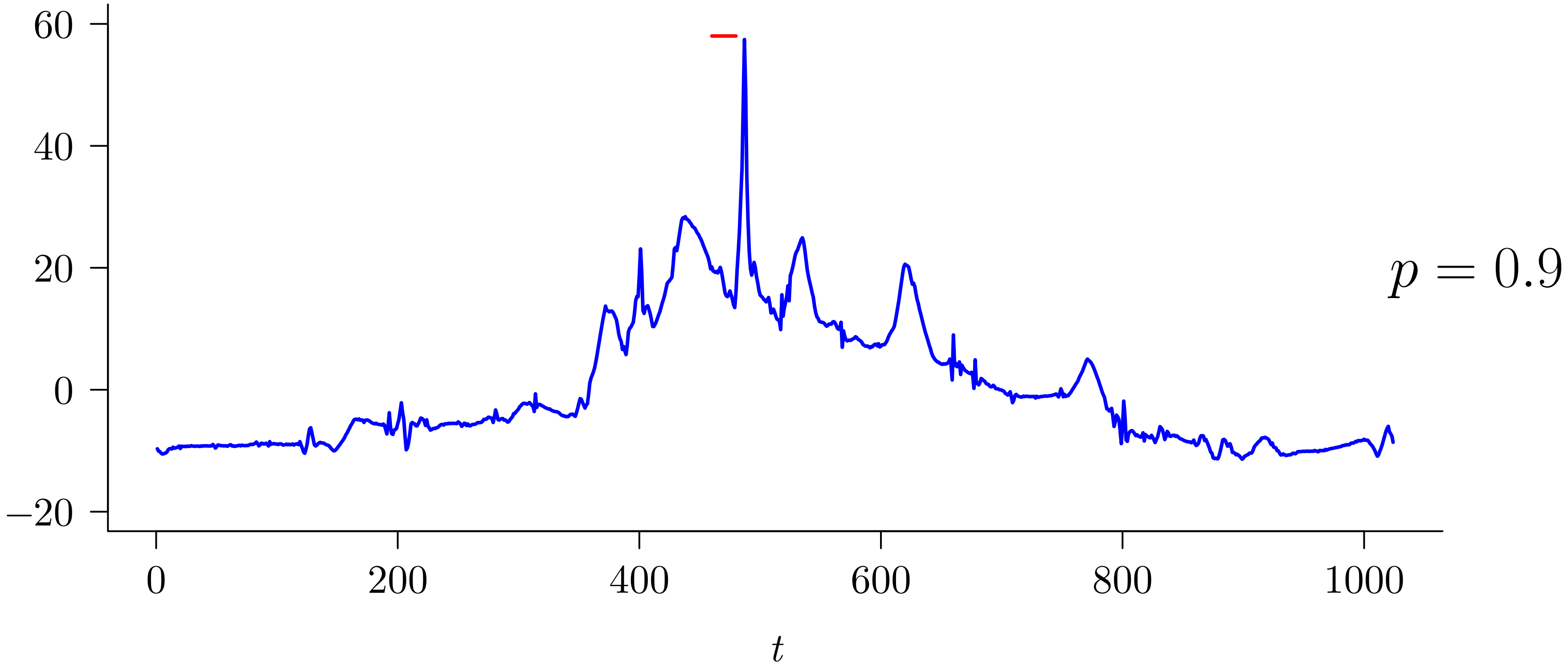
- same as before, but now with $p = 0.5$

Examples of Wavelet-Based Shrinkage: IV



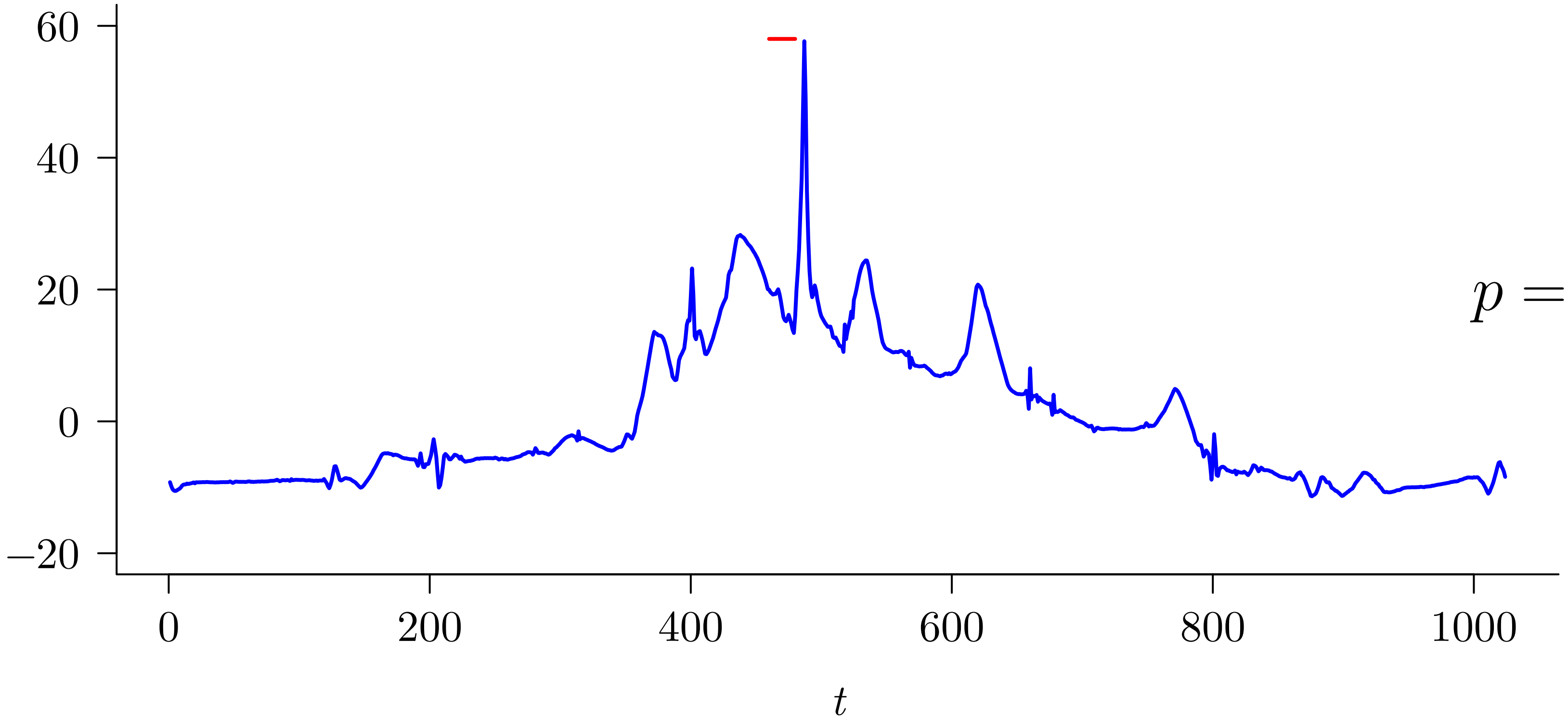
- same as before, but now with $p = 0.75$

Examples of Wavelet-Based Shrinkage: V



- same as before, but now with $p = 0.9$

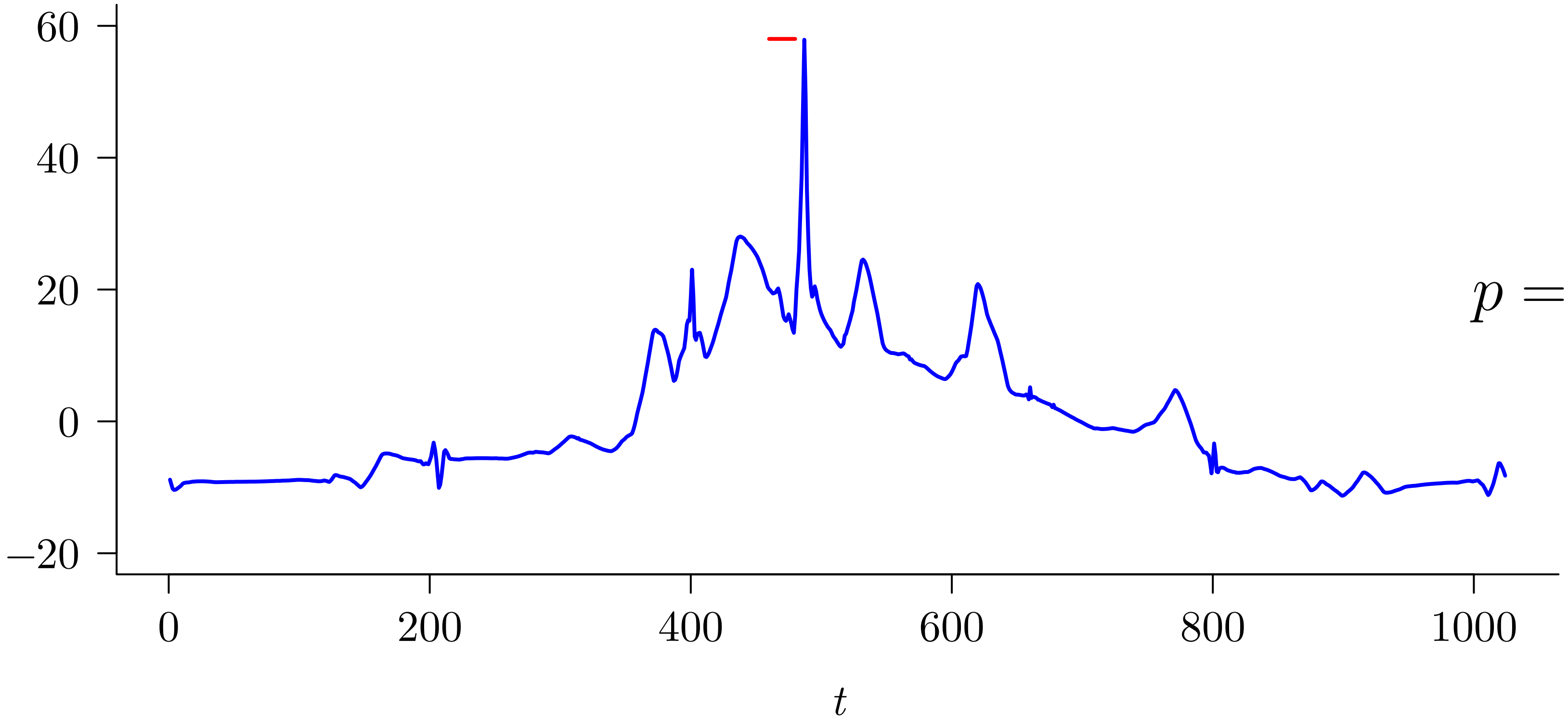
Examples of Wavelet-Based Shrinkage: VI



$p = 0.95$

- same as before, but now with $p = 0.95$

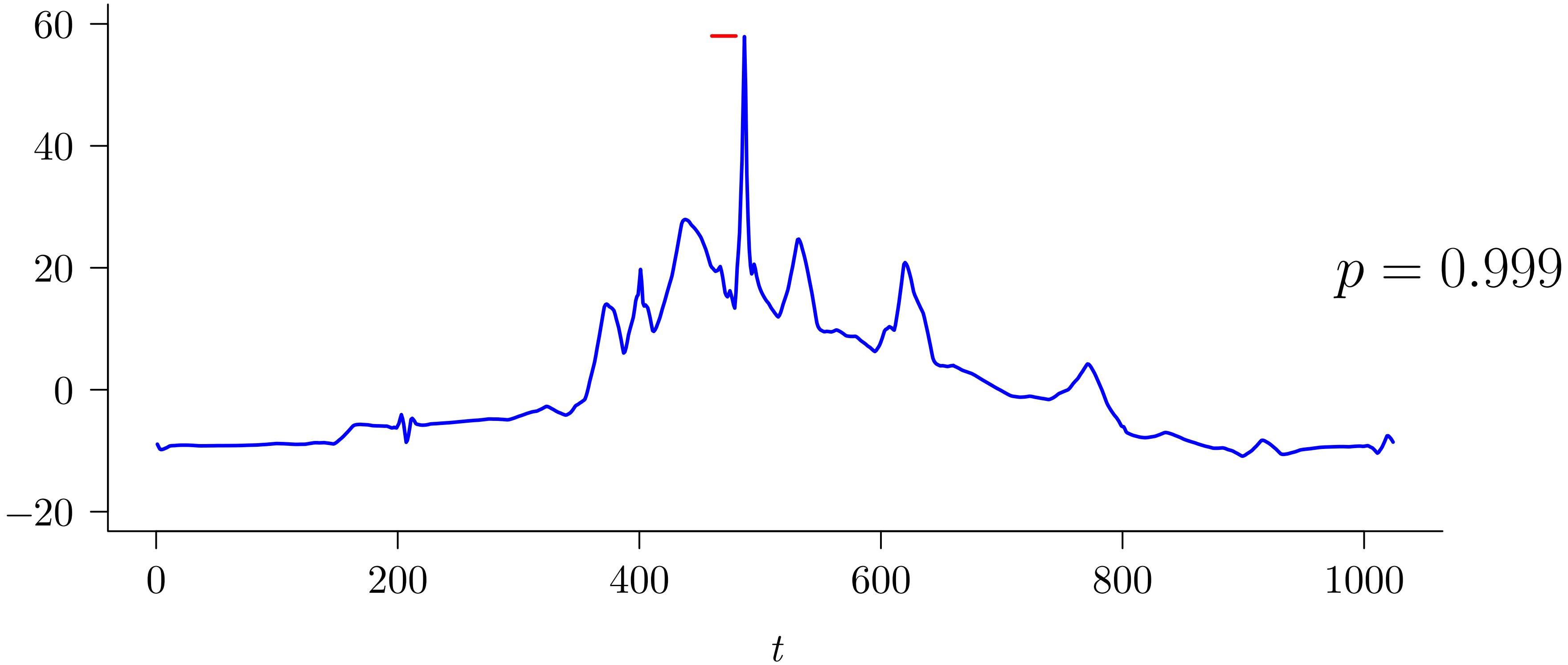
Examples of Wavelet-Based Shrinkage: VII



$p = 0.99$

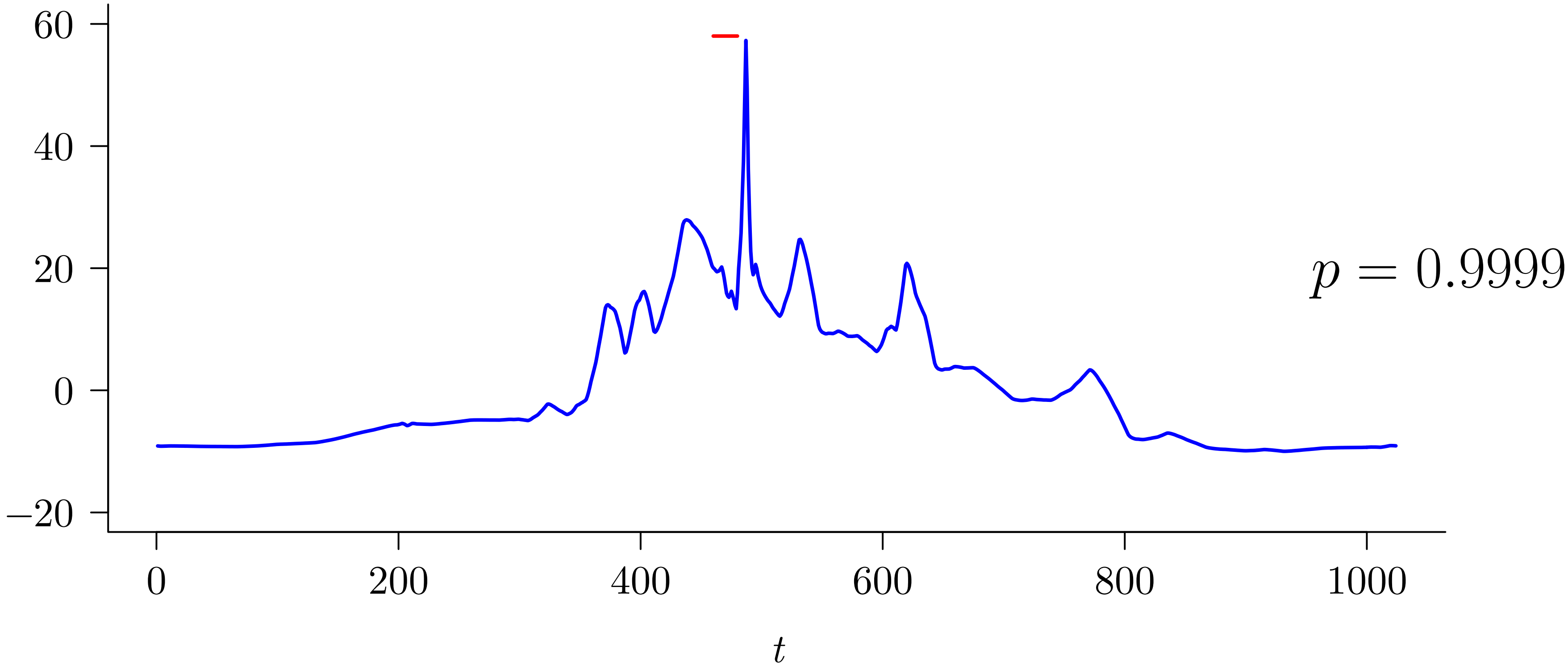
- same as before, but now with $p = 0.99$

Examples of Wavelet-Based Shrinkage: VIII



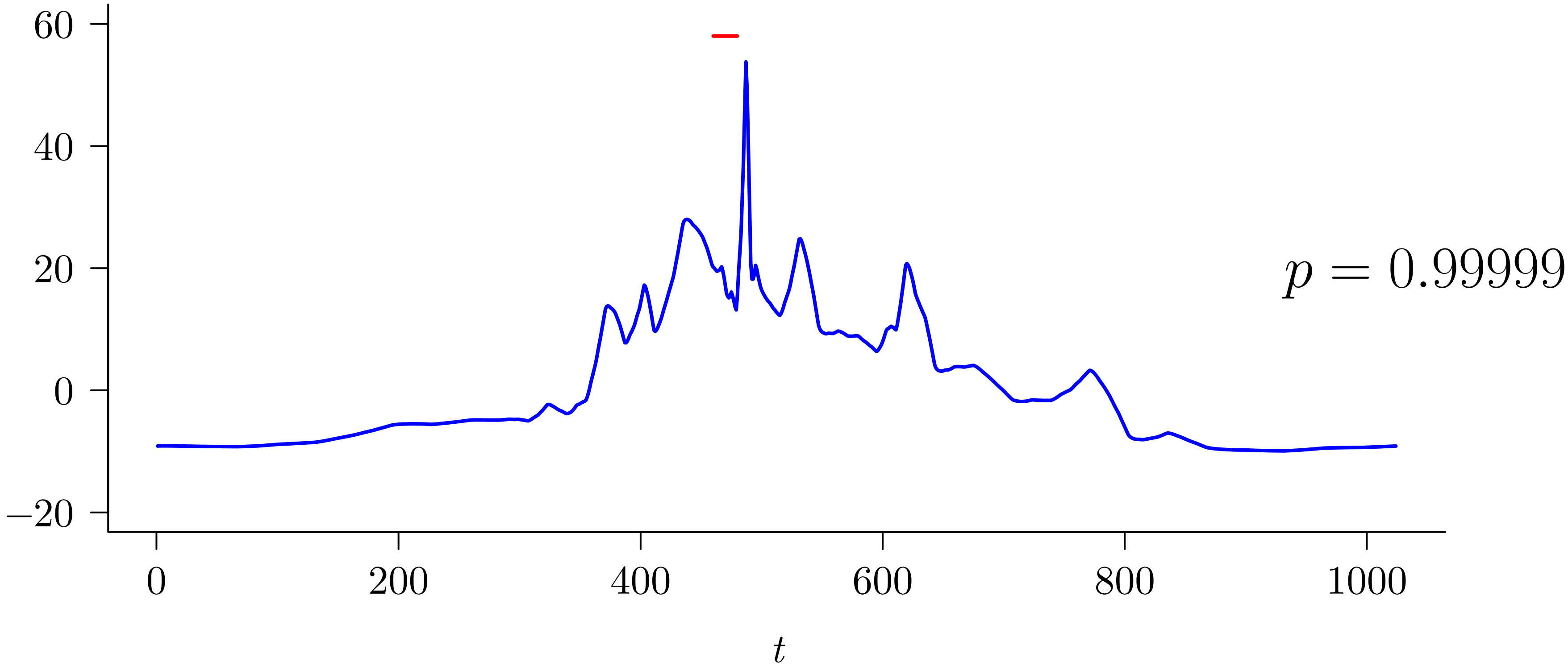
- same as before, but now with $p = 0.999$

Examples of Wavelet-Based Shrinkage: IX



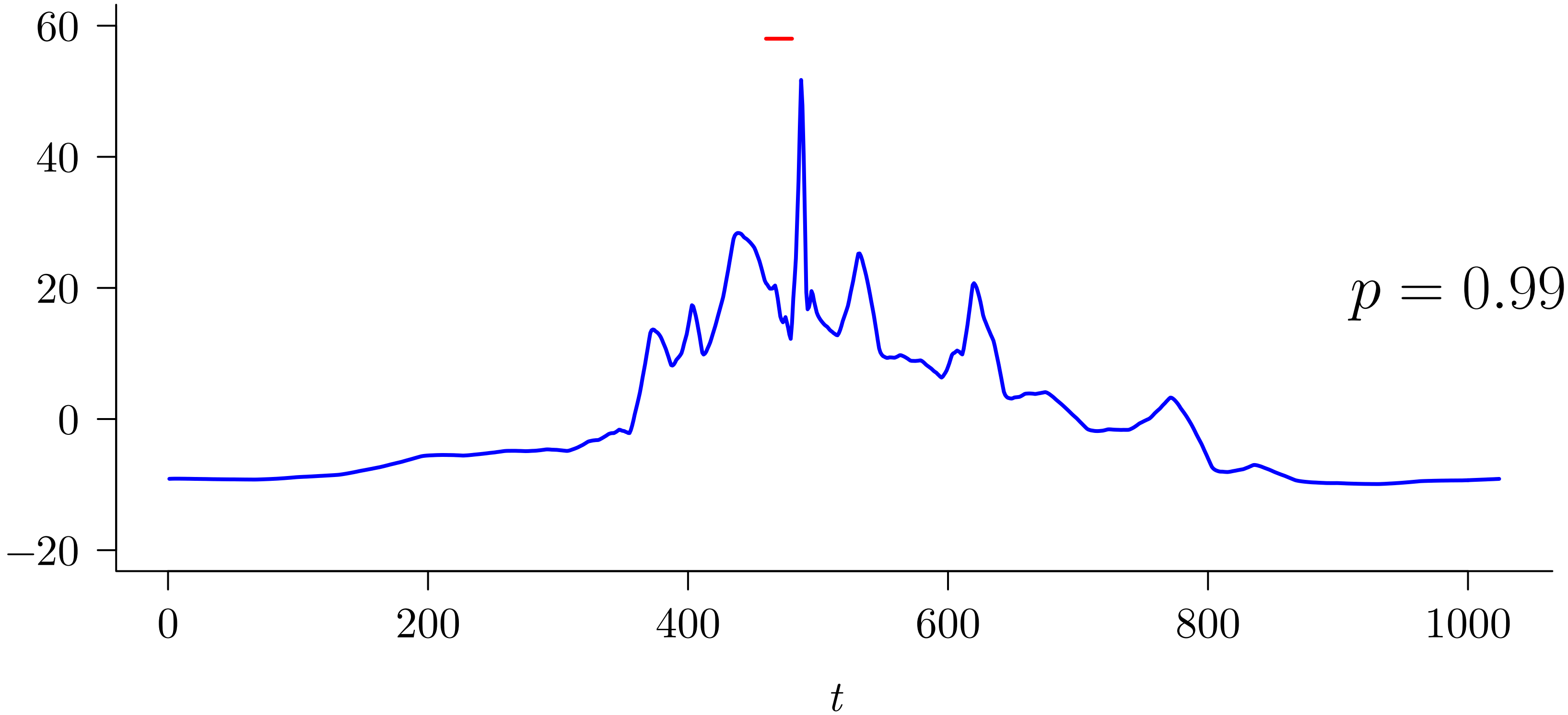
- same as before, but now with $p = 0.9999$

Examples of Wavelet-Based Shrinkage: X



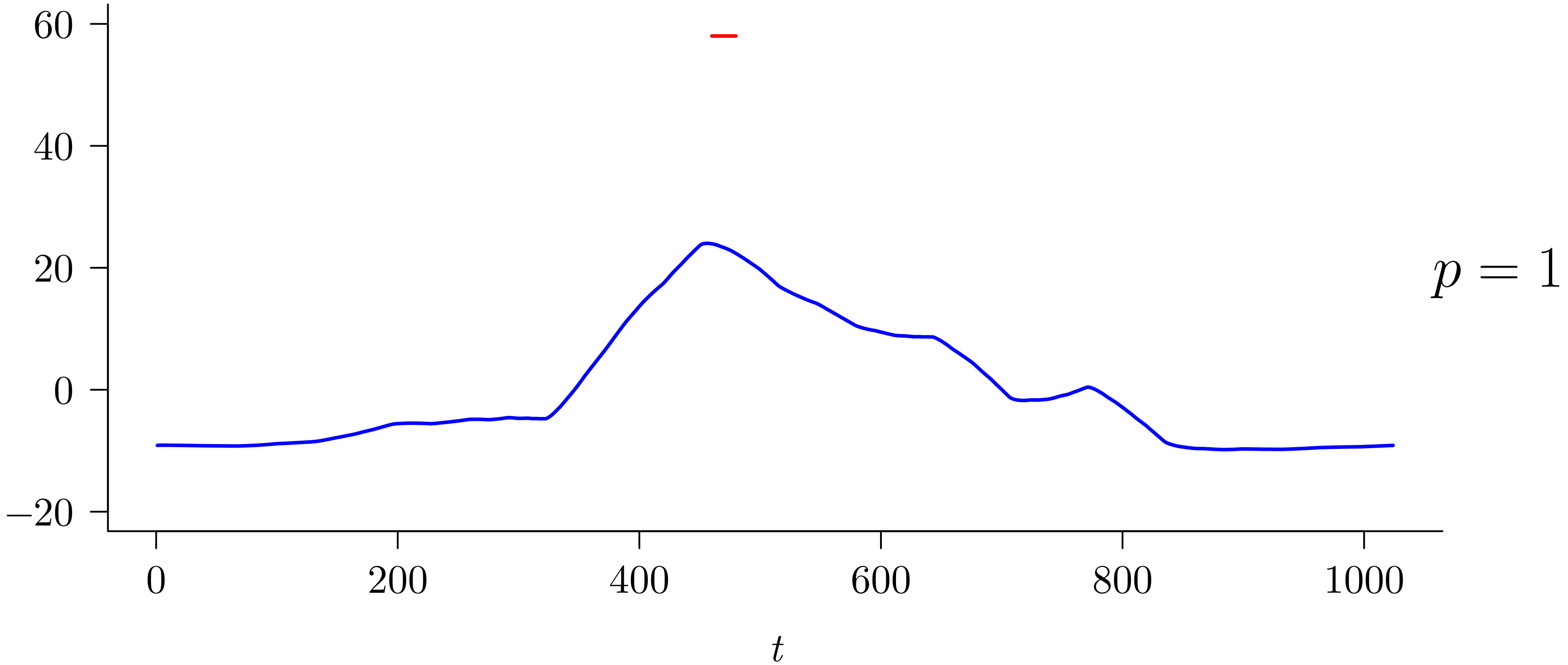
- same as before, but now with $p = 0.99999$

Examples of Wavelet-Based Shrinkage: XI



- same as before, but now with $p = 0.999999$

Examples of Wavelet-Based Shrinkage: XII



- same as before, but now with $p = 1$

Shrinkage Functions

- conditional mean estimator takes form

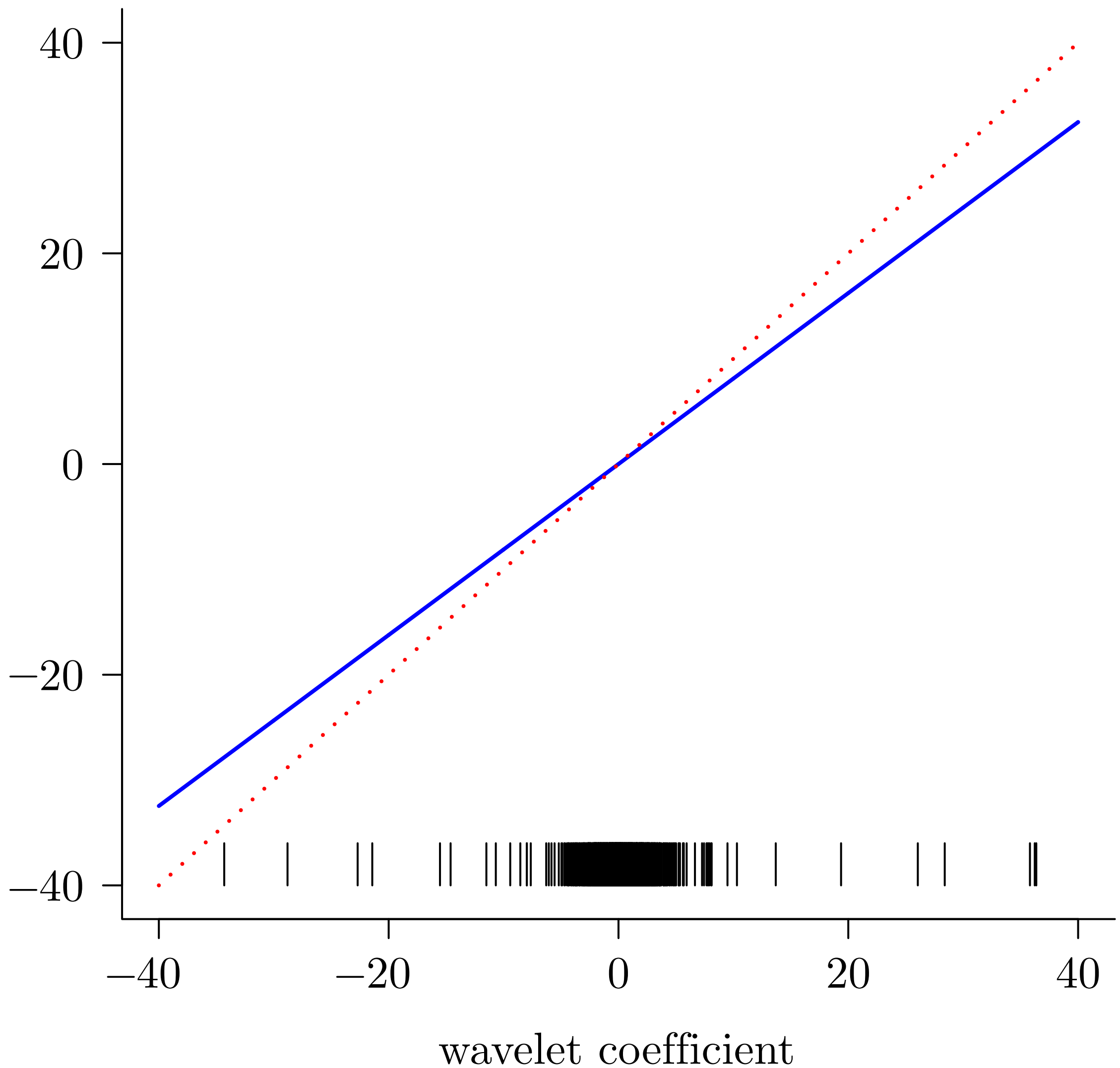
$$E\{R_{j,t} \mid W_{j,t}\} = \frac{b}{1 + c_{j,t}} W_{j,t},$$

where

$$b \equiv \frac{\sigma_G^2}{\sigma_G^2 + \sigma_\epsilon^2} \;\; \text{and} \;\; c_{j,t} = \frac{p\surd(\sigma_G^2 + \sigma_\epsilon^2)}{(1-p)\sigma_\epsilon} e^{-bW_{j,t}^2/(2\sigma_\epsilon^2)}$$

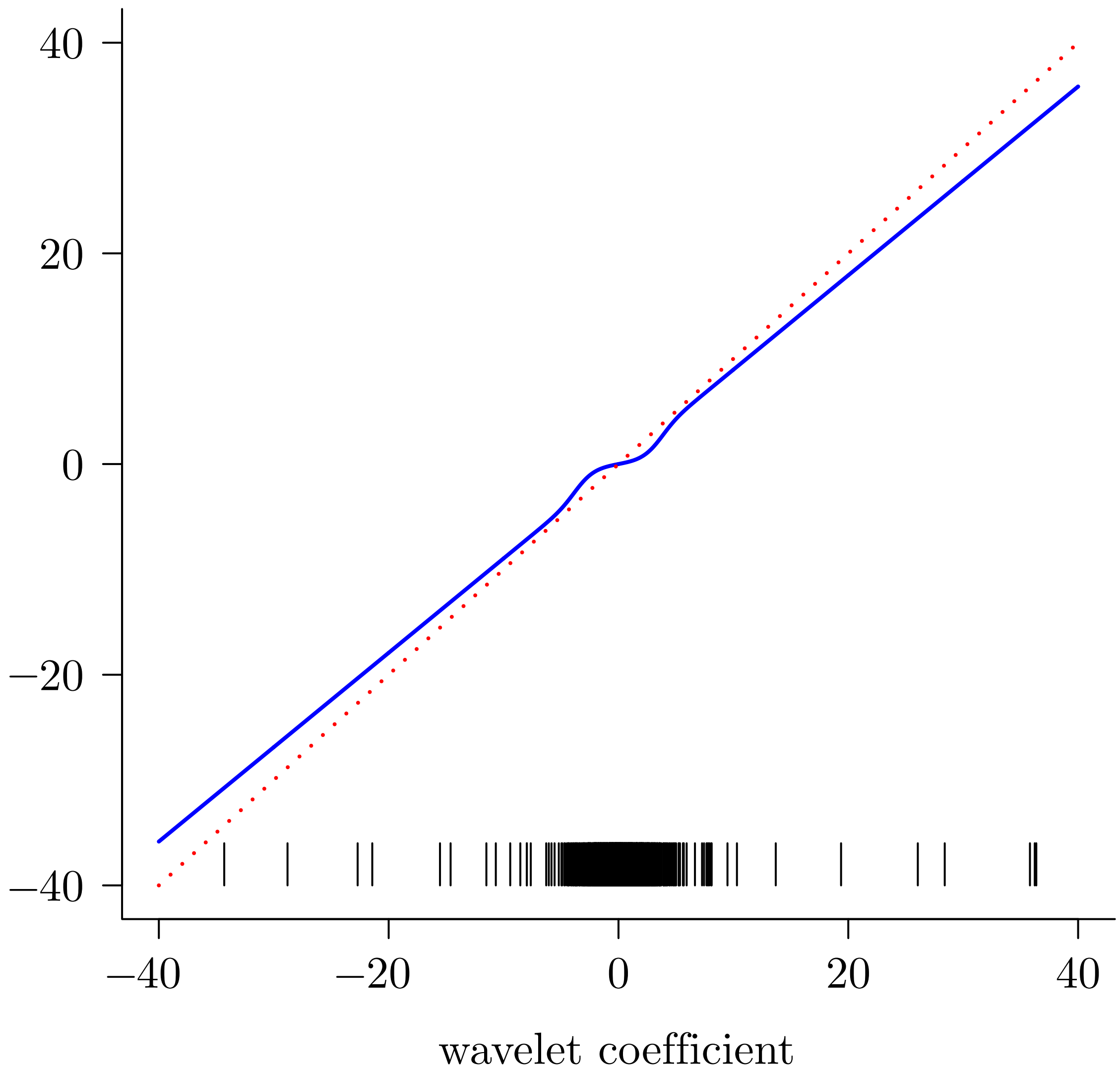
- shrinkage function determined once σ_ϵ^2, σ_G^2 and p are set
- following plots show shrinkage function $\frac{b}{1+c_{j,t}}W_{j,t}$ versus $W_{j,t}$ for various selections of p as $W_{j,t}$ ranges from -40 to 40
- note: actual $W_{j,t}$'s for NMR series range from -34.3 to 36.4, with values indicated by short vertical lines at bottom of plots

Examples of Shrinkage Functions: I



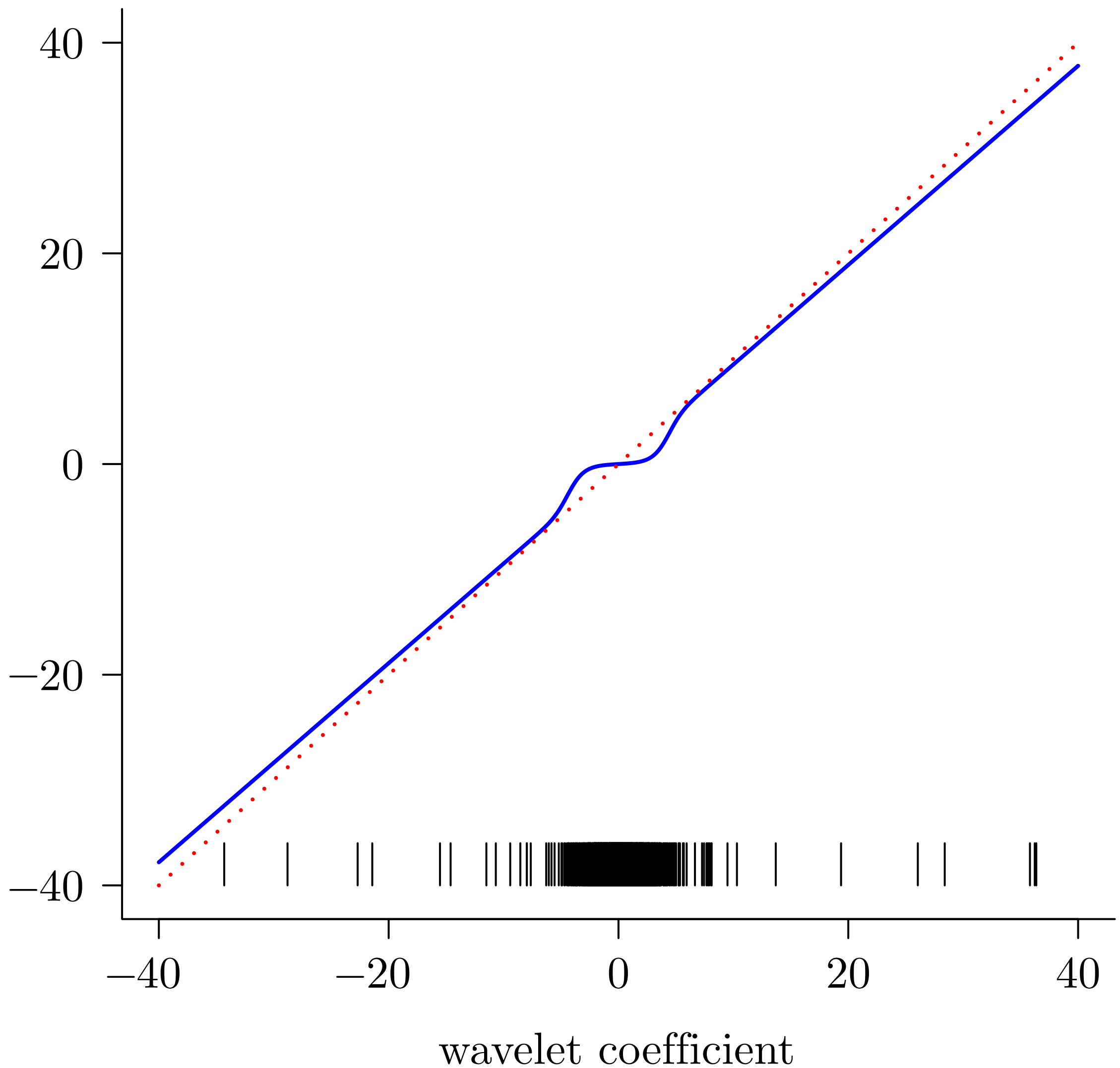
$p = 0$

Examples of Shrinkage Functions: II



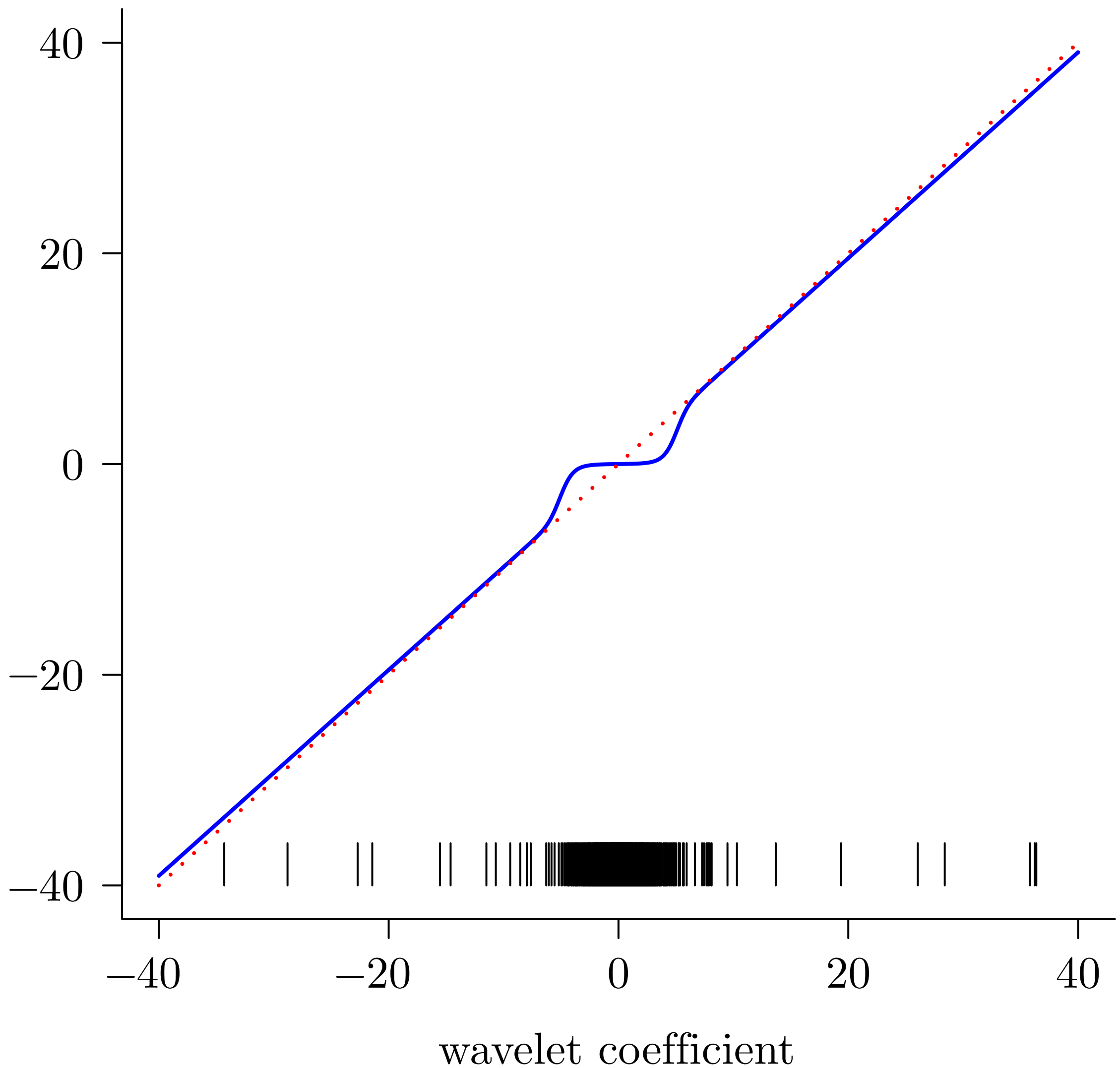
$p = 0.5$

Examples of Shrinkage Functions: III



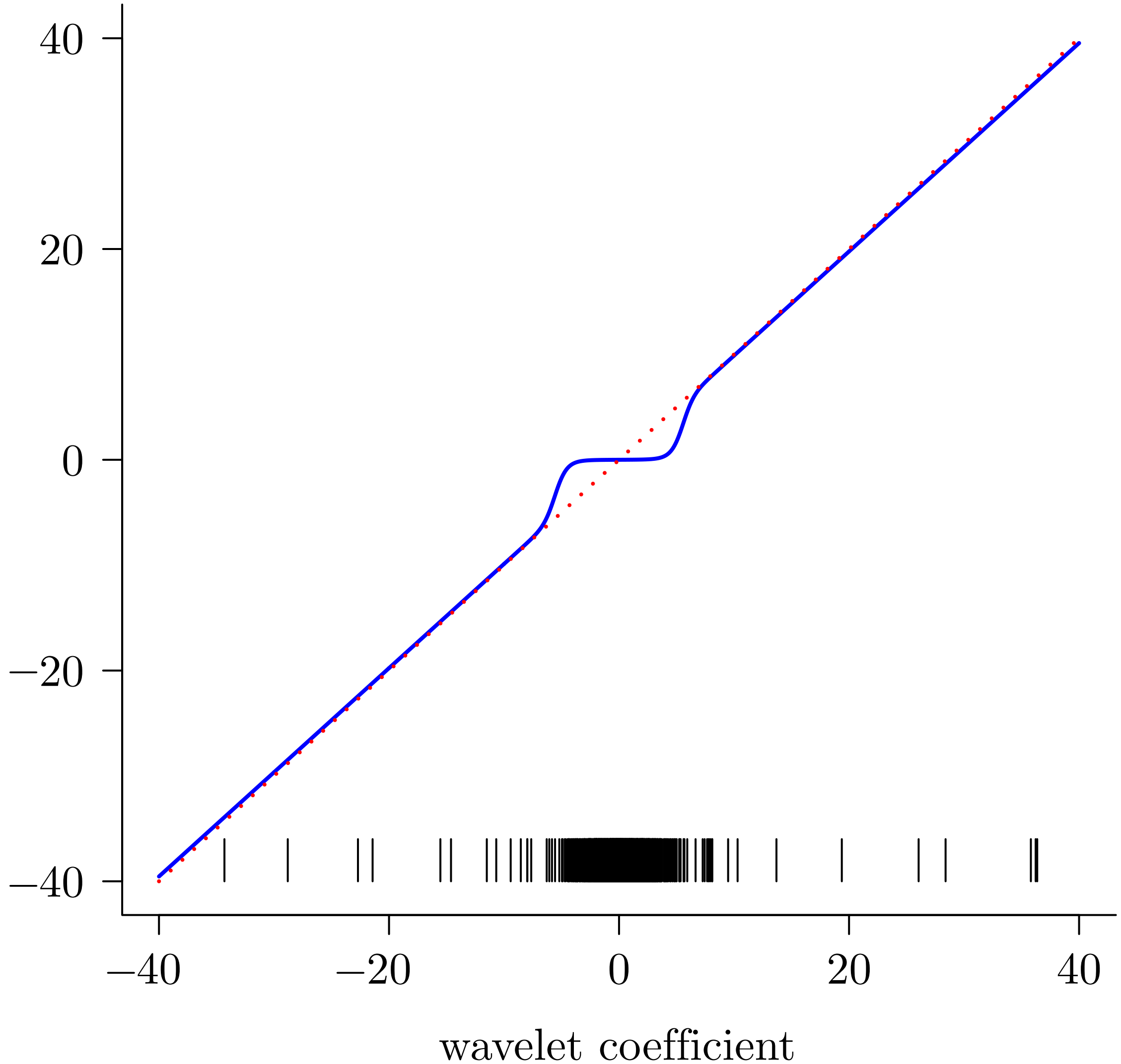
$p = 0.75$

Examples of Shrinkage Functions: IV



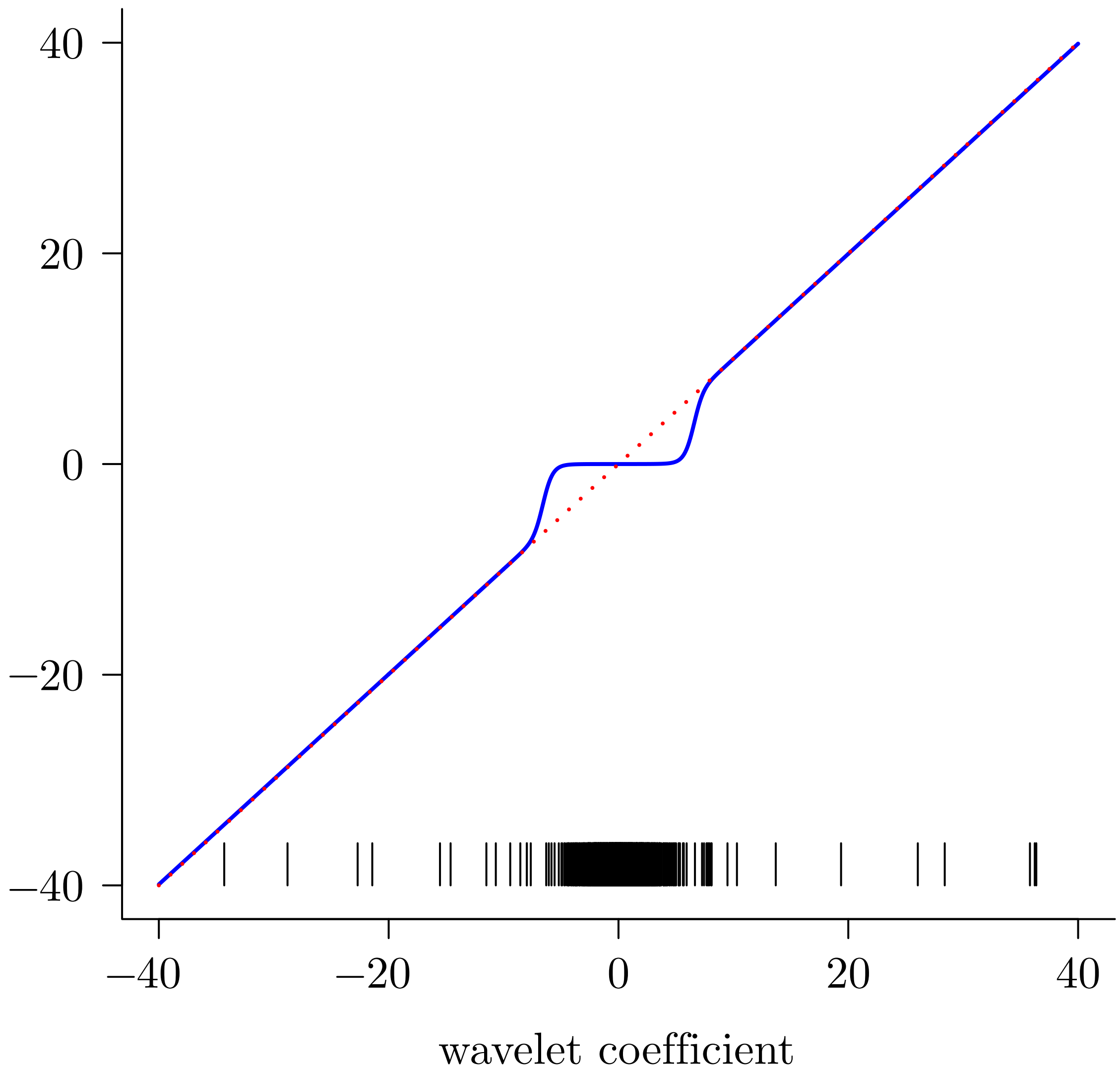
$p = 0.9$

Examples of Shrinkage Functions: V



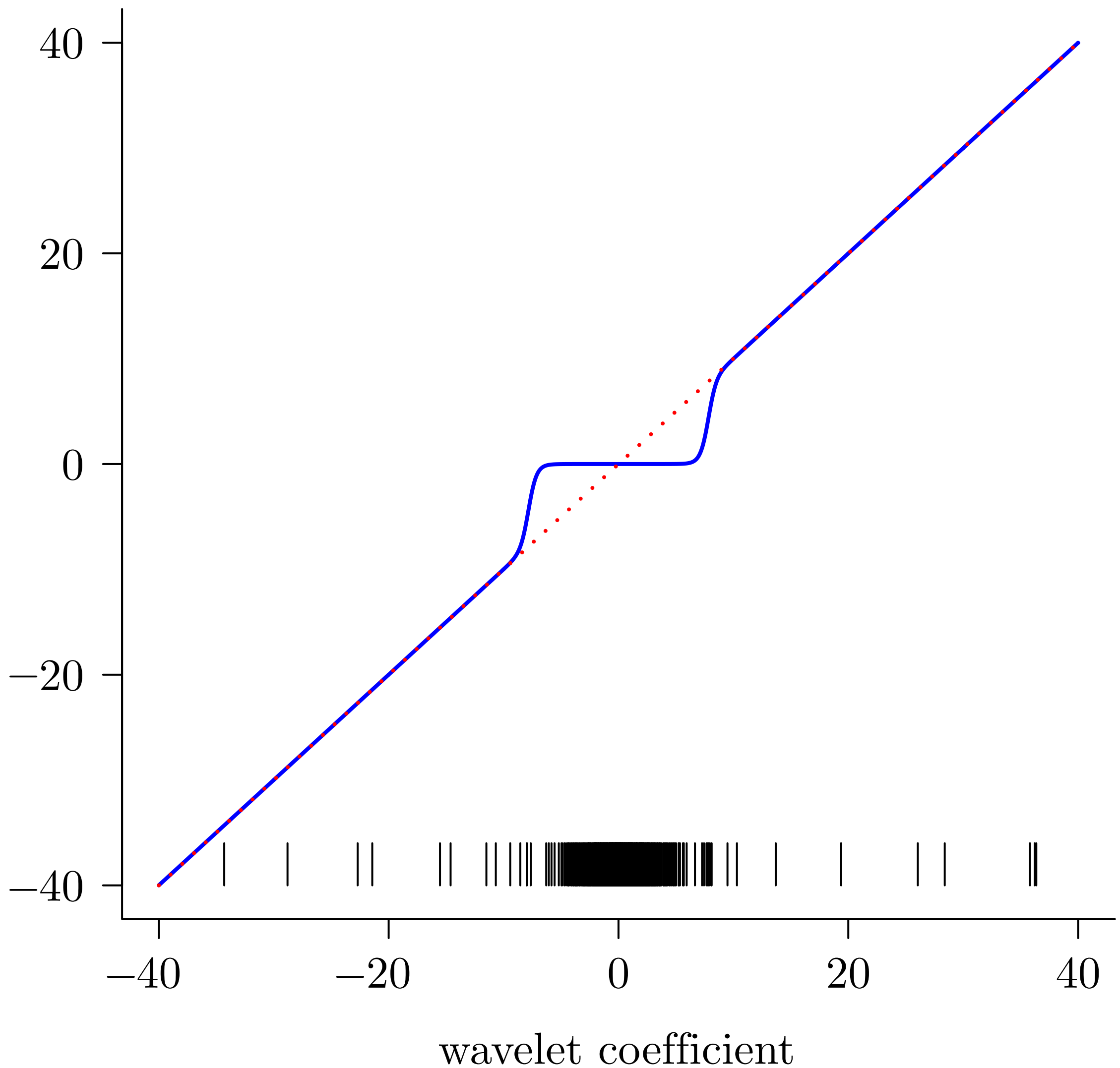
$p = 0.95$

Examples of Shrinkage Functions: VI



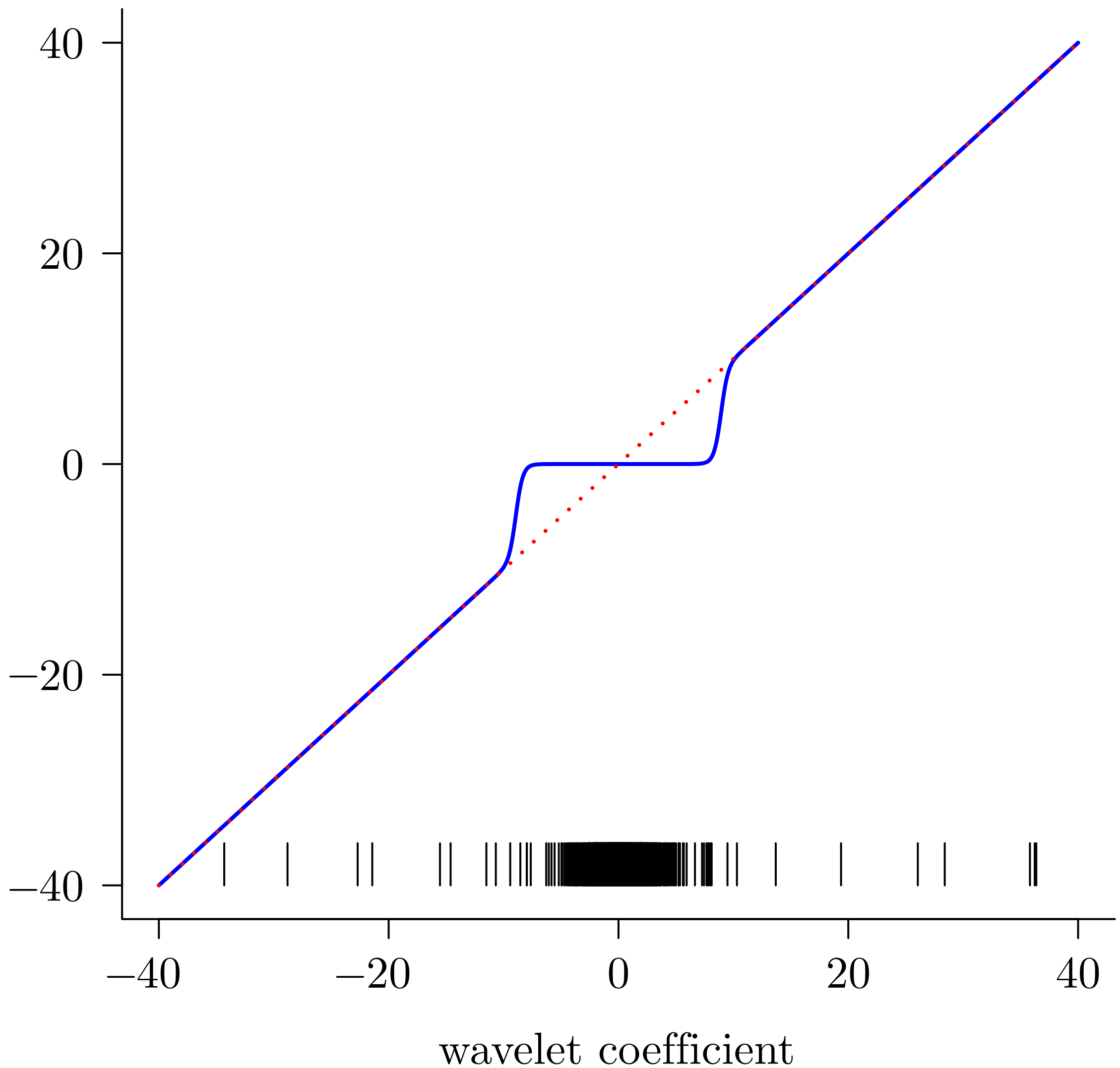
$p = 0.99$

Examples of Shrinkage Functions: VII



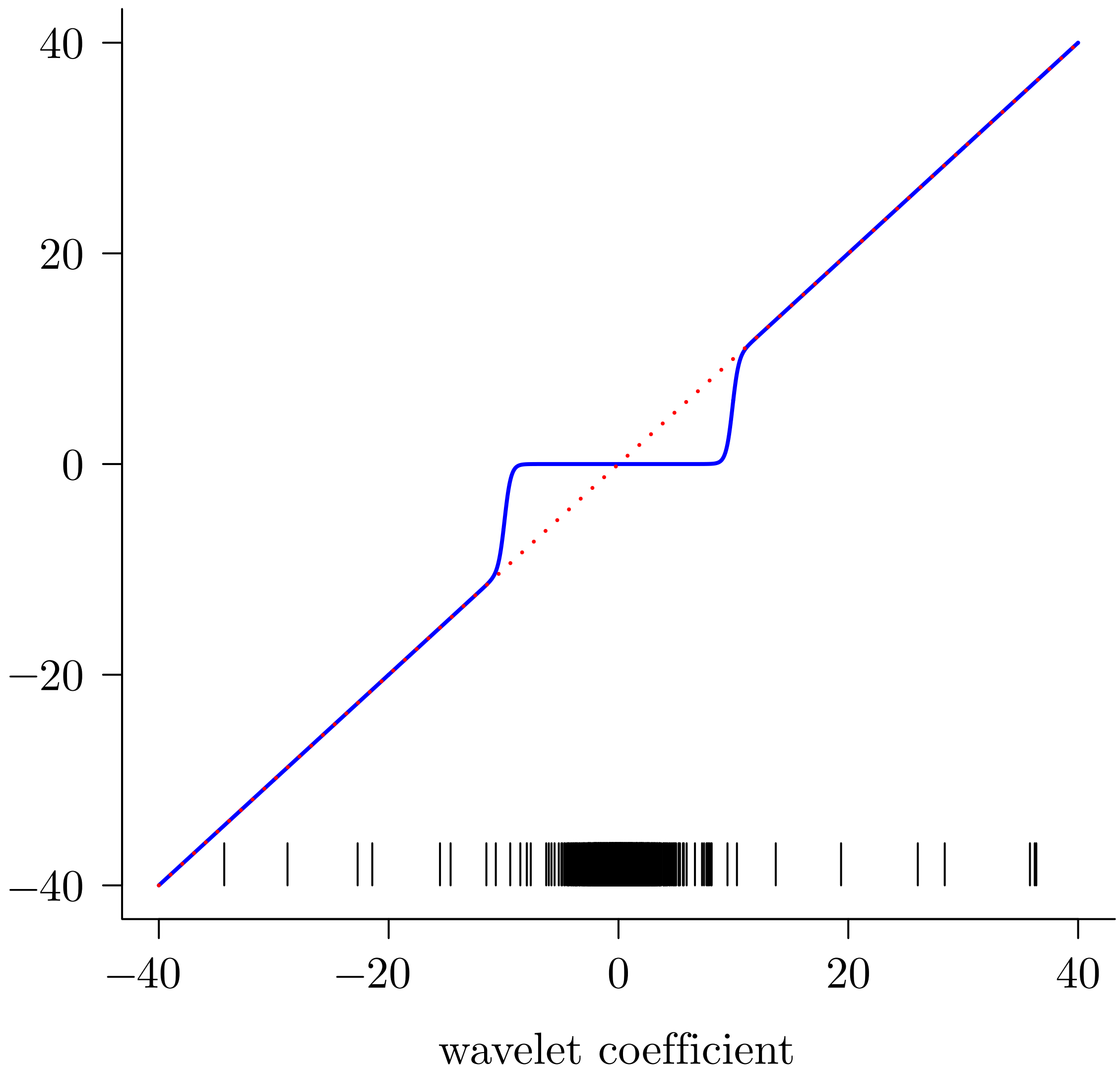
$p = 0.999$

Examples of Shrinkage Functions: VIII



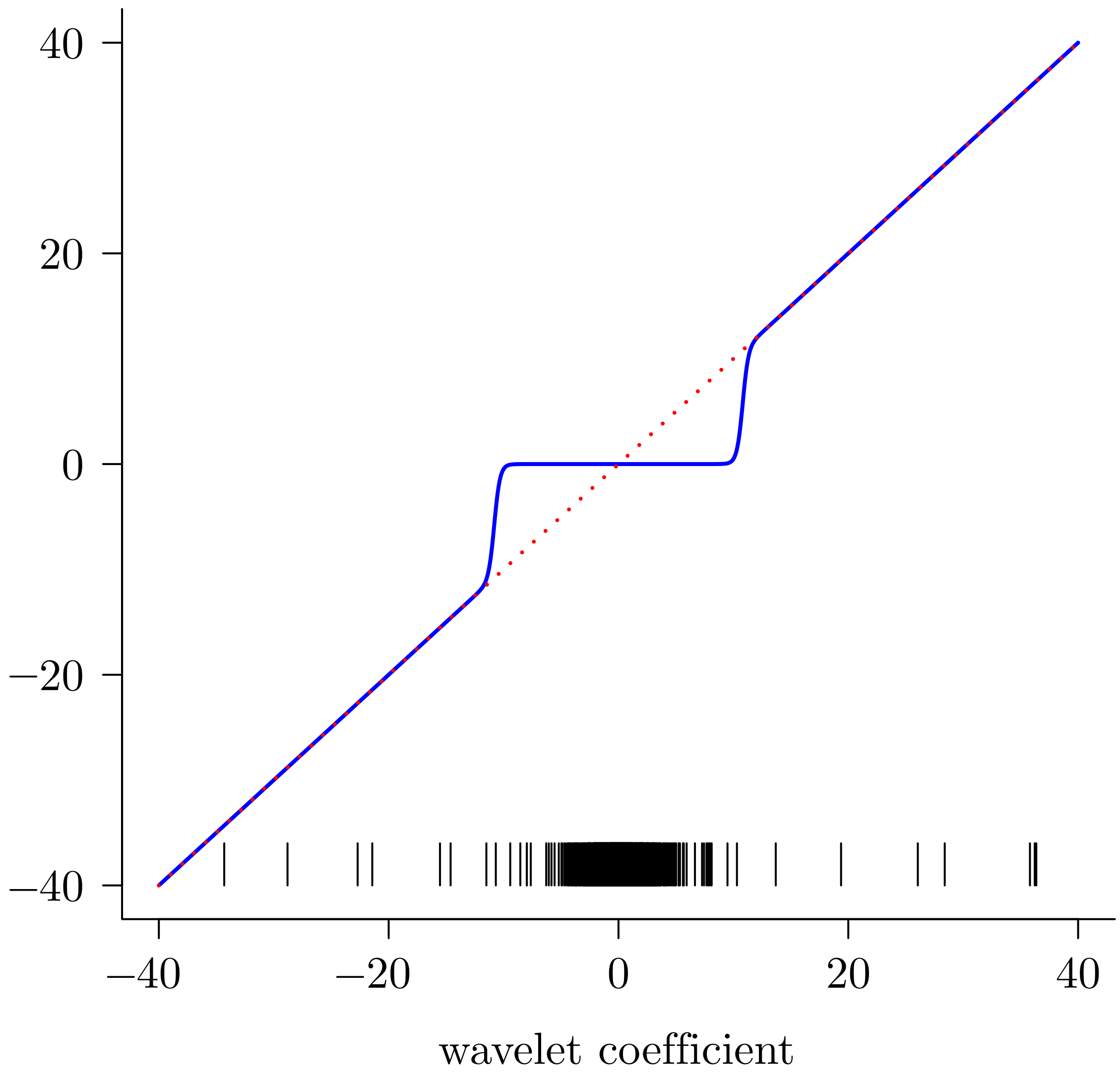
$p = 0.9999$

Examples of Shrinkage Functions: IX



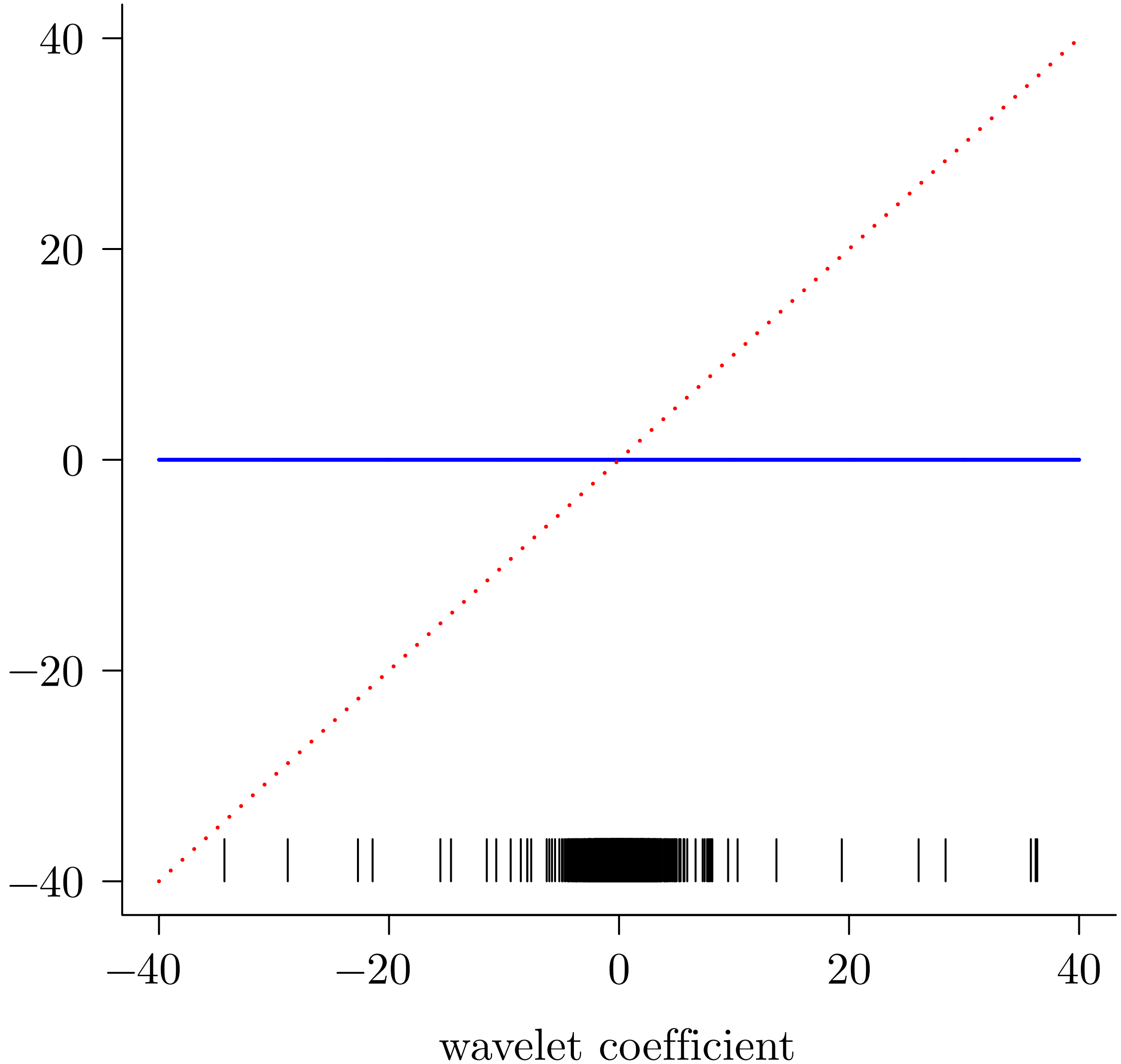
$p = 0.99999$

Examples of Shrinkage Functions: X



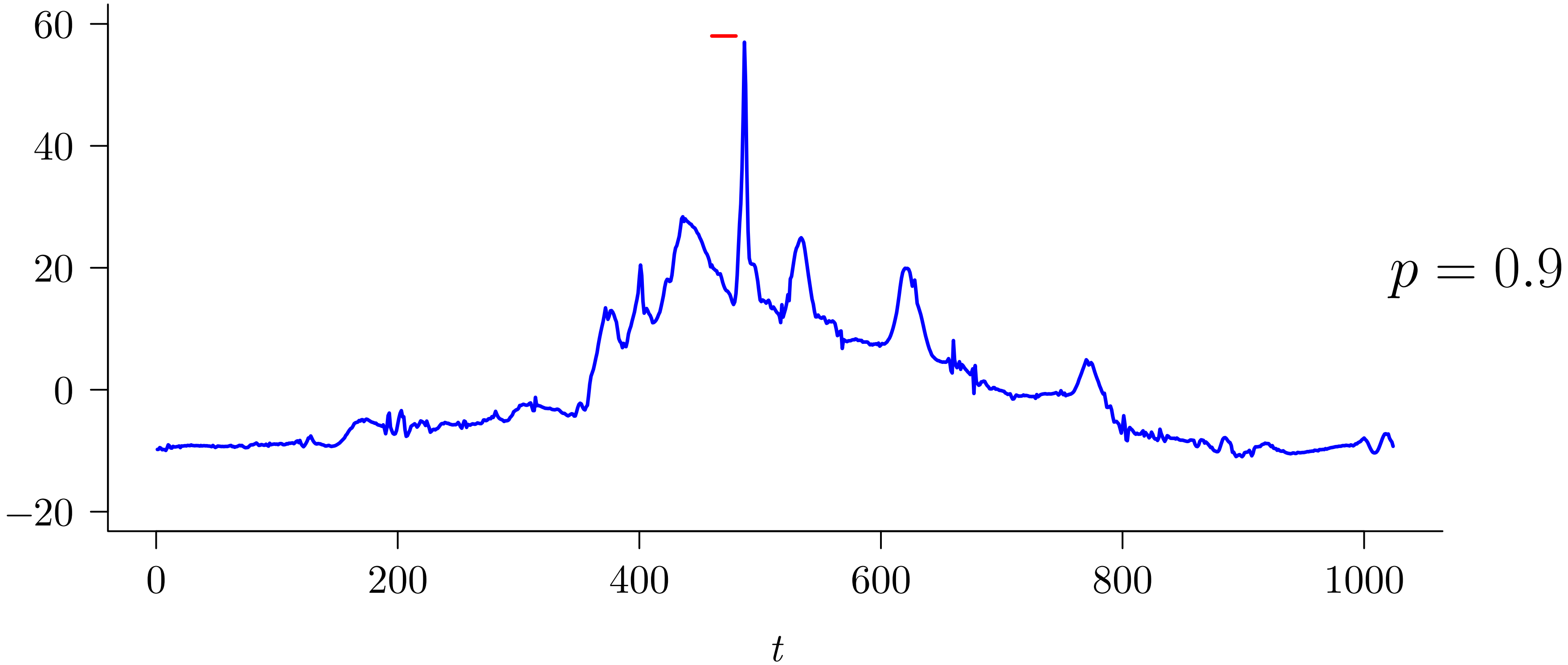
$p = 0.999999$

Examples of Shrinkage Functions: XI



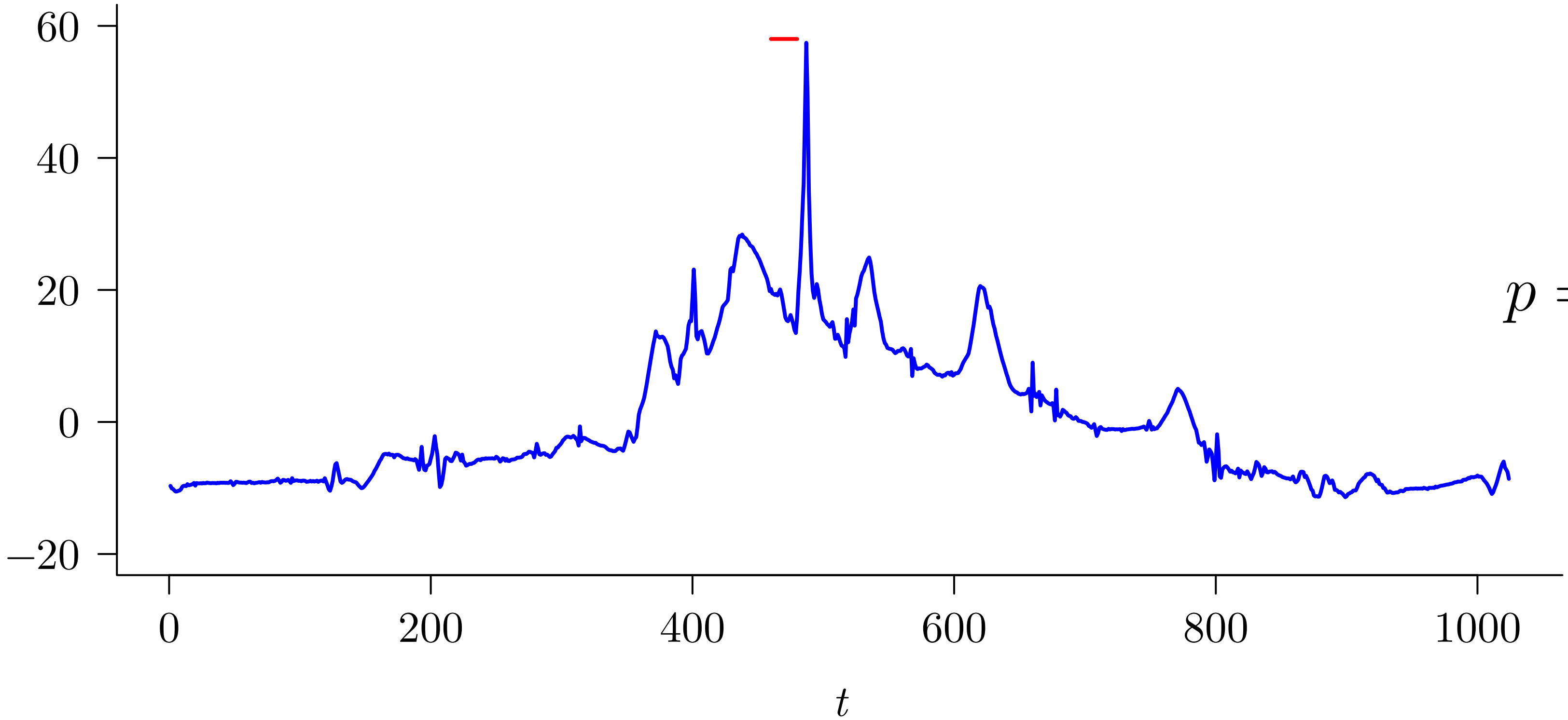
$p = 1$

Wavelet-Based Shrinkage with Cycle Spinning: I



- same as before, but now with $p = 0.9$ and cycle spinning

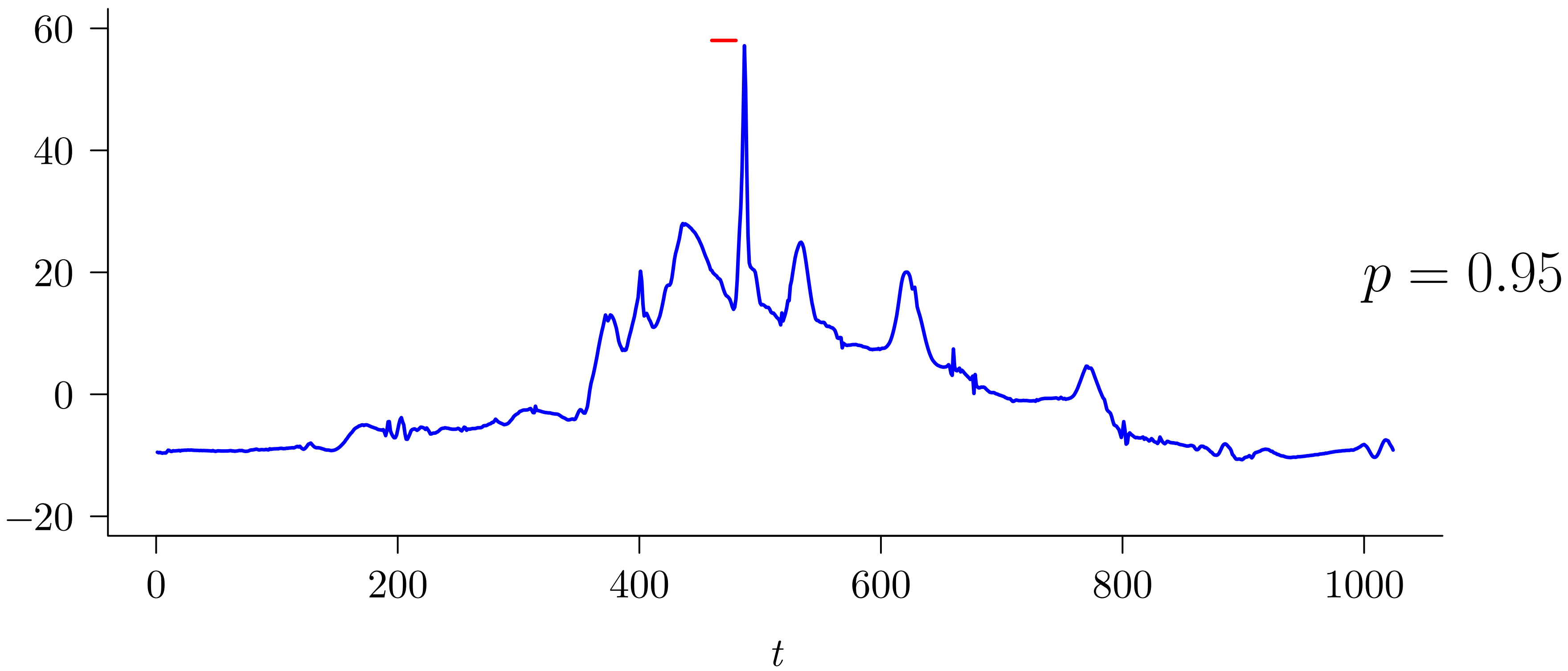
Examples of Wavelet-Based Shrinkage: V



$p = 0.9$

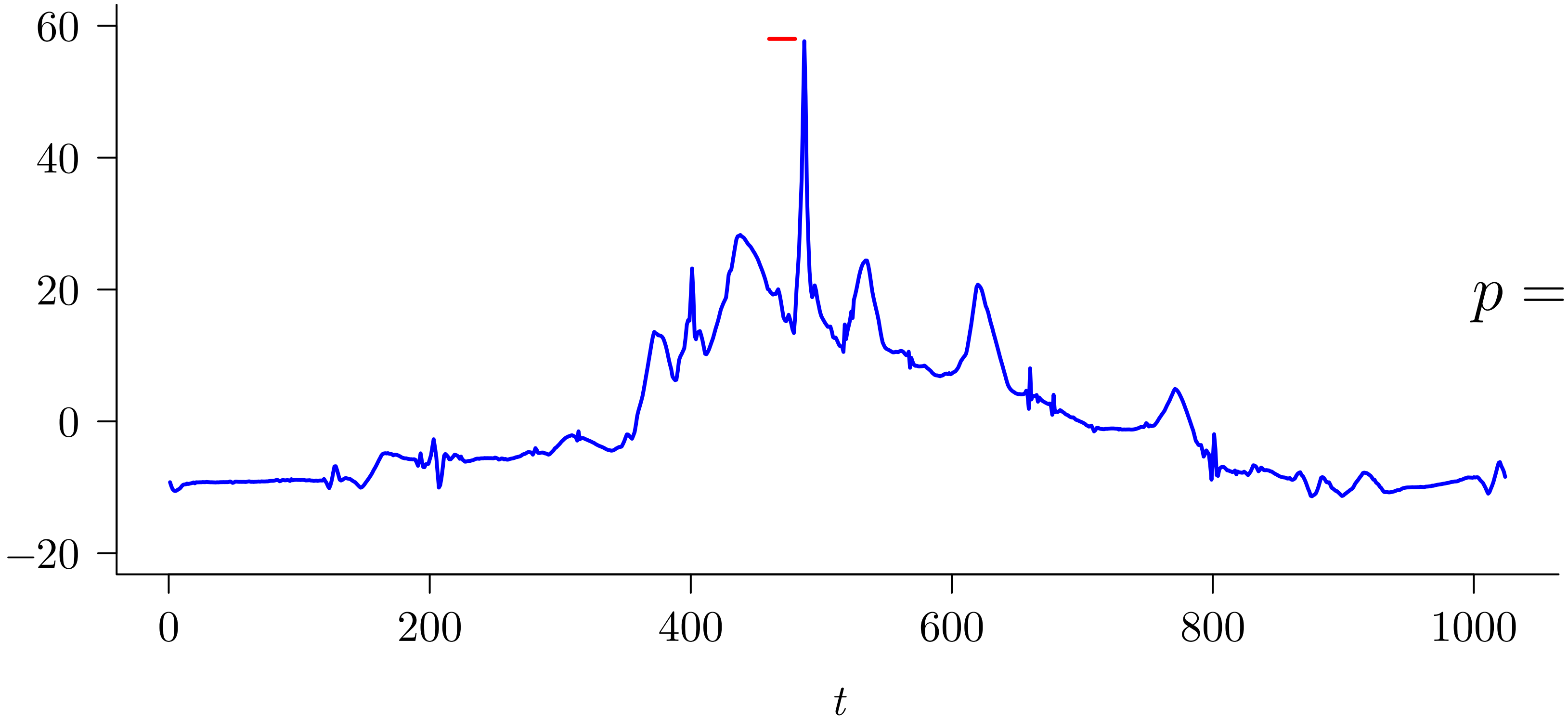
- same as before, but now with $p = 0.9$

Wavelet-Based Shrinkage with Cycle Spinning: II



- same as before, but now with $p = 0.95$ and cycle spinning

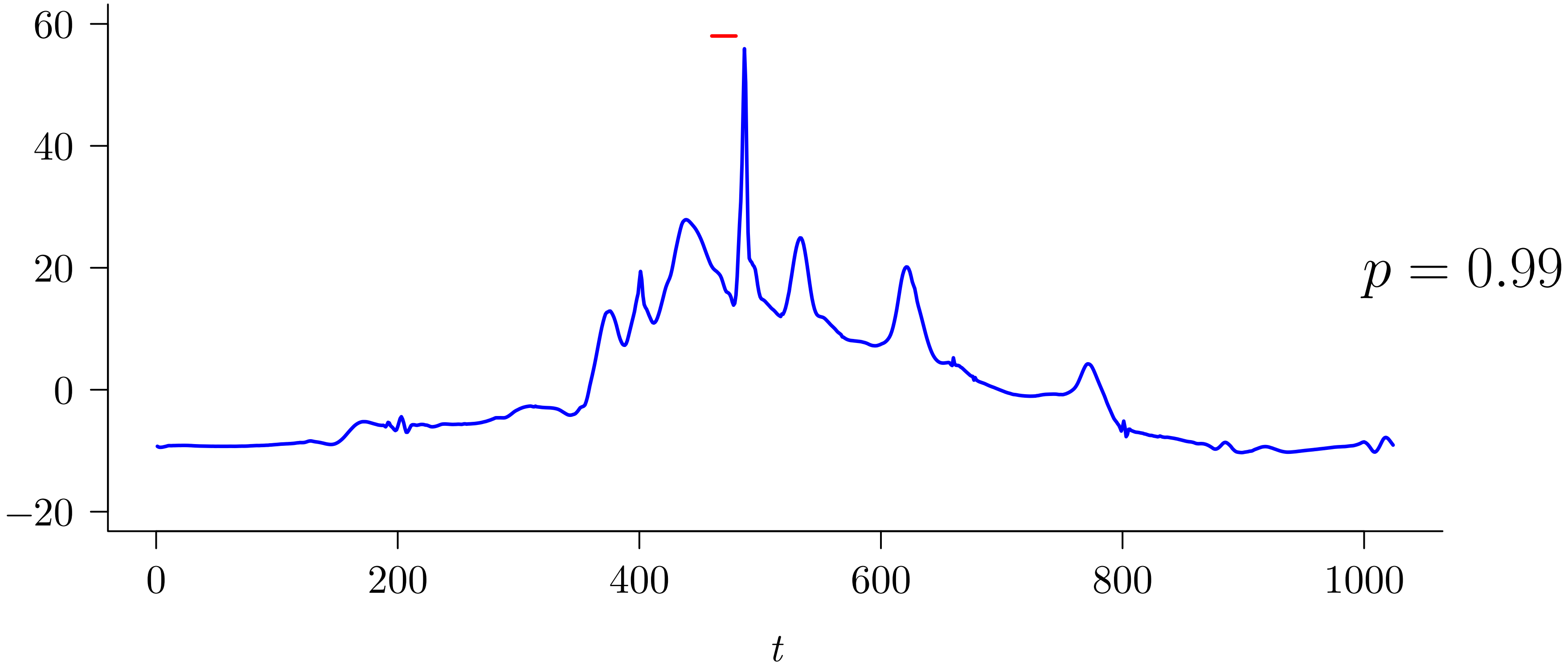
Examples of Wavelet-Based Shrinkage: VI



$p = 0.95$

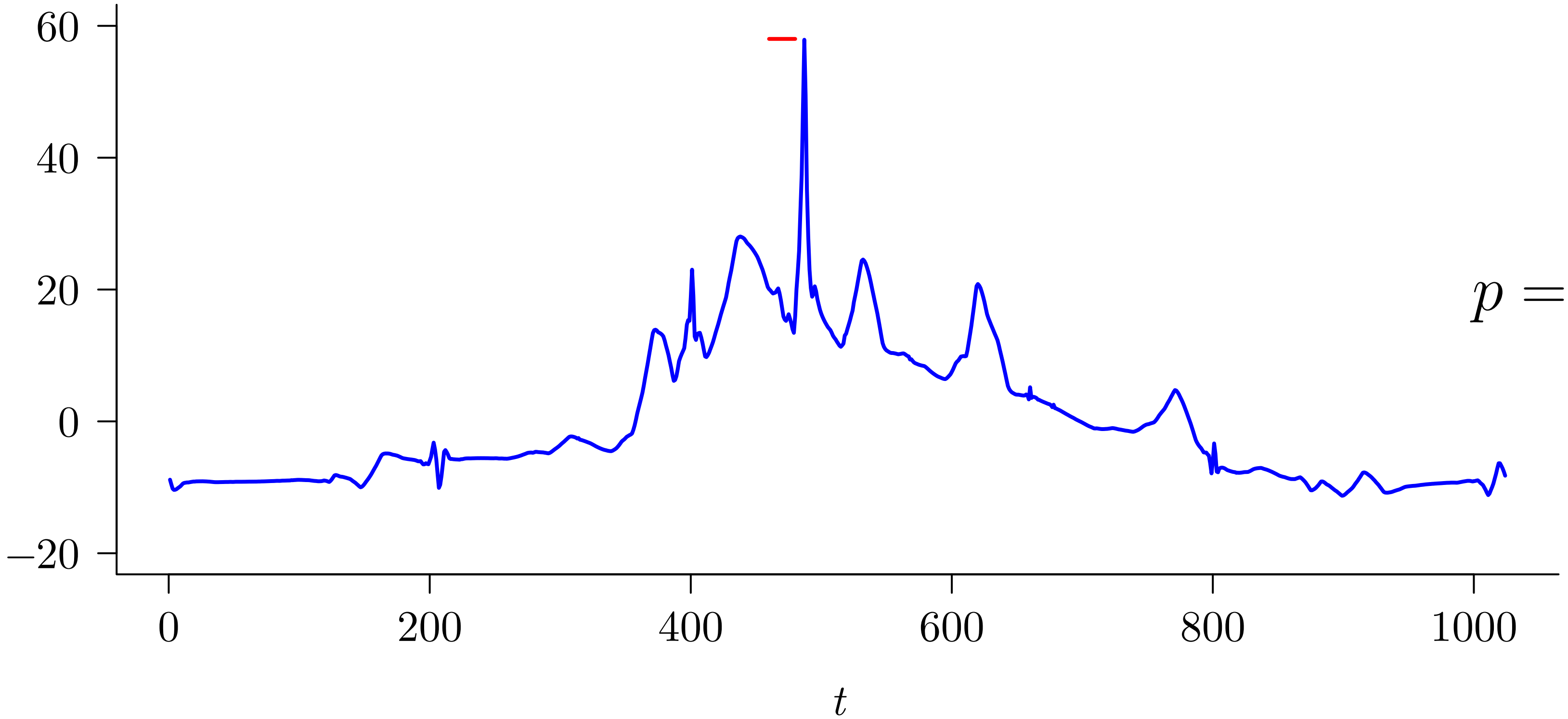
- same as before, but now with $p = 0.95$

Wavelet-Based Shrinkage with Cycle Spinning: III



- same as before, but now with $p = 0.99$ and cycle spinning

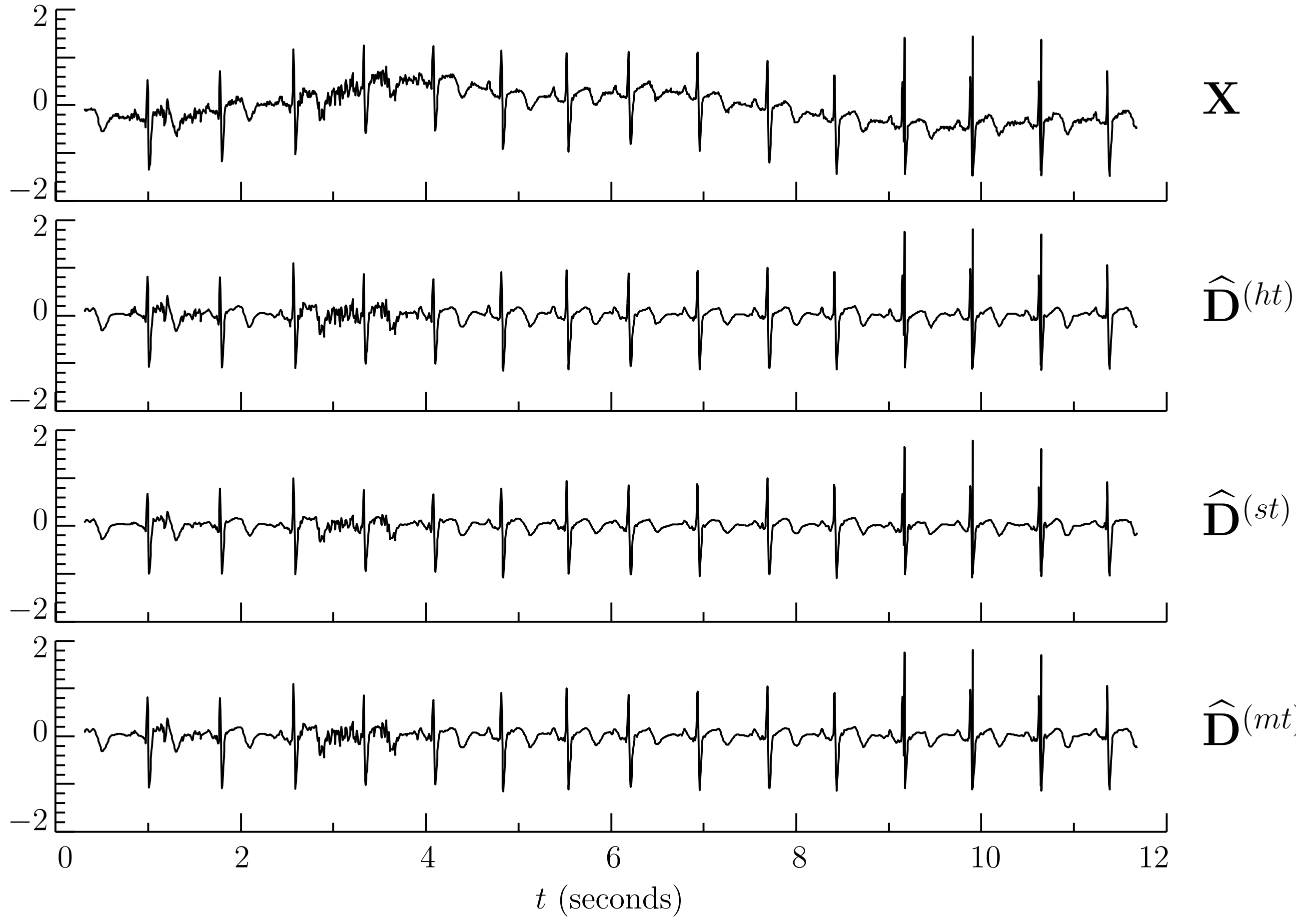
Examples of Wavelet-Based Shrinkage: VII



$p = 0.99$

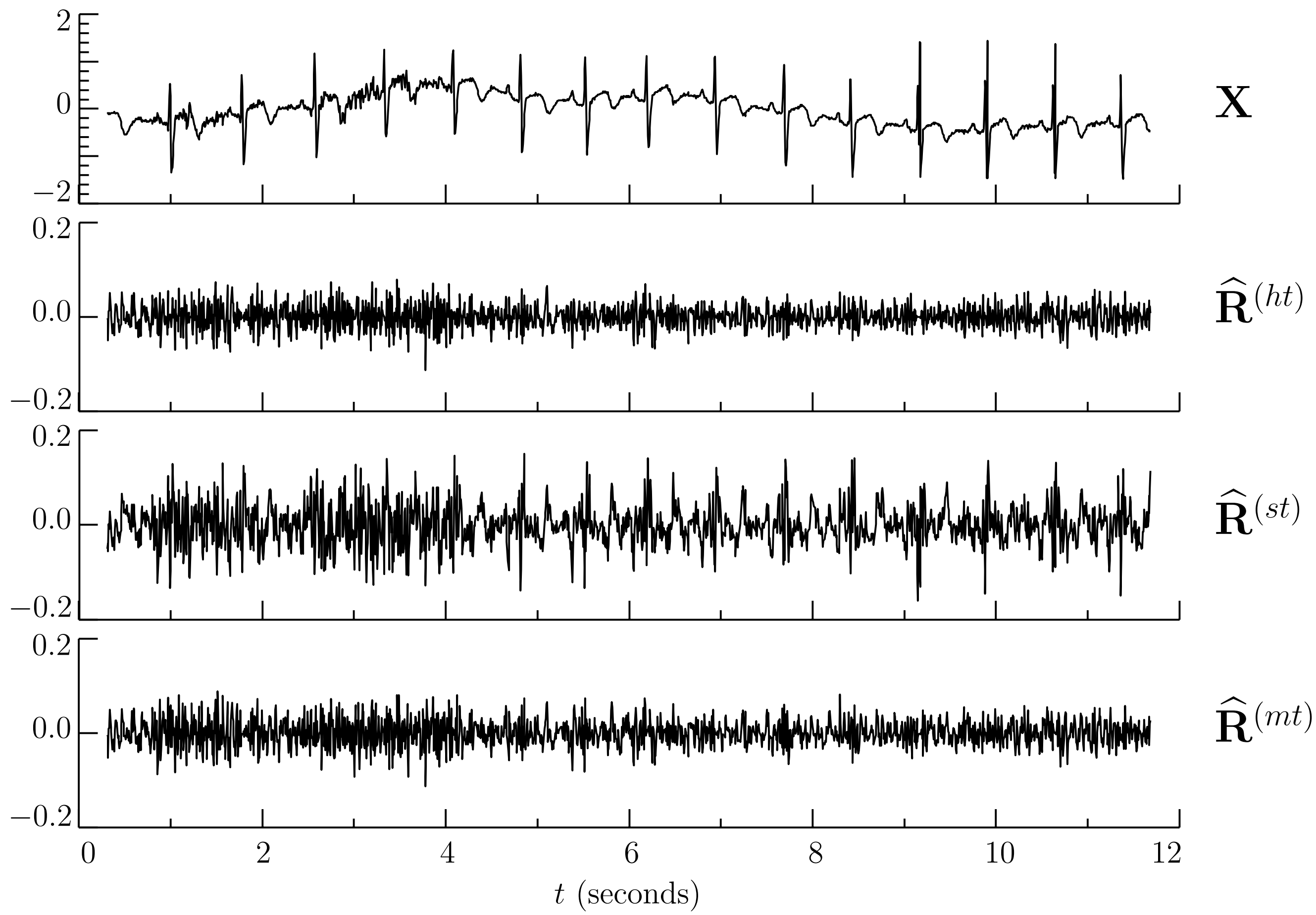
- same as before, but now with $p = 0.99$

Case Study – Denoising ECG Time Series: I



- hard/soft/mid threshold estimates with $J_0 = 6$ partial LA(8) DWT, MAD & scaling coefficients to 0 (zaps baseline drift)

Case Study – Denoising ECG Time Series: II



- residuals from signal estimates, i.e., $\widehat{\mathbf{R}}^{(t)} = \mathbf{X} - \widehat{\mathbf{D}}^{(t)}$ (assumption of constant noise variance is questionable)

SDF Estimation via Periodogram: I

- let $\{X_t\}$ be a stationary process with mean 0 and variance σ_X^2
- spectral density function (SDF) $S(\cdot)$ describes $\{X_t\}$ by decomposing σ_X^2 on a frequency by frequency basis:

$$\int_{-1/2}^{1/2} S(f)\, df = \sigma_X^2$$

- suppose we observe a time series that is a realization of a portion $X_0, \ldots, X_{N-1}$ of $\{X_t\}$, and we want to form a consistent estimator $\hat{S}(f)$ of $S(f)$; i.e., want

$$E\{\hat{S}(f)\} \to S(f) \text{ and } \operatorname{var}\{\hat{S}(f)\} \to 0 \text{ as } N \to \infty$$

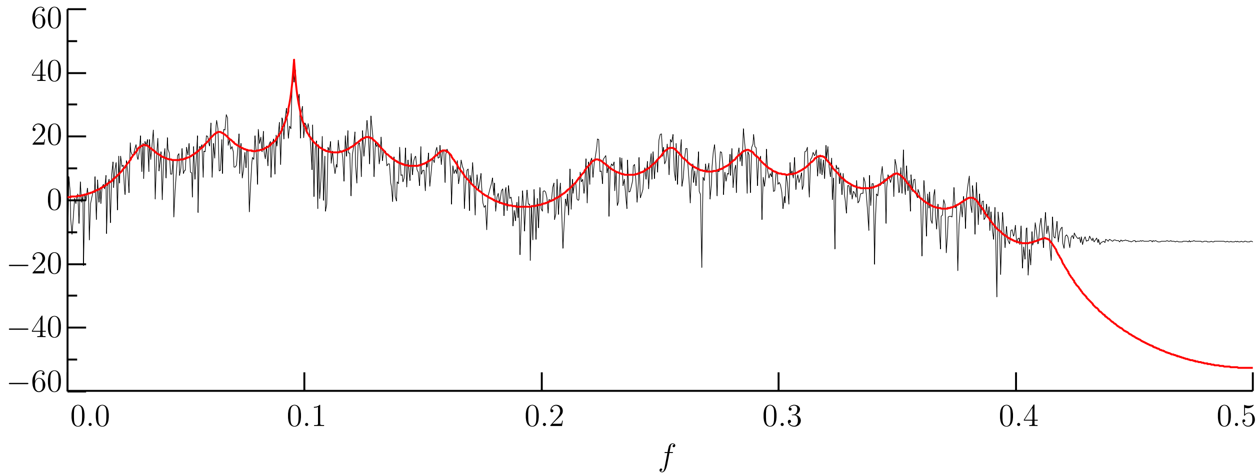
SDF Estimation via Periodogram: II

- the most basic estimator of $S(f)$ is the periodogram:

$$\hat{S}^{(p)}(f) \equiv \frac{1}{N} \left| \sum_{t=0}^{N-1} X_t e^{-i2\pi f t} \right|^2, \qquad |f| \leq 1/2$$

- for large N and $0 < f < 1/2$, statistical theory says that $\hat{S}^{(p)}(f)$ has a distribution given by $S(f)\chi_2^2/2$, where χ_2^2 is a chi-square RV with 2 degrees of freedom
- if N is large enough (might need to be *very* large!), have
  - $E\{\hat{S}^{(p)}(f)\} \approx E\{S(f)\chi_2^2/2\} = S(f)$
  - $\text{var}\,\{\hat{S}^{(p)}(f)\} \approx \text{var}\,\{S(f)\chi_2^2/2\} = S^2(f)$
- conclusion: as $N \to \infty$, $\text{var}\,\{\hat{S}^{(p)}(f)\} \to S^2(f) \neq 0$ in general; i.e., periodogram is an inconsistent estimator of $S(f)$

Example of SDF and Periodogram



- periodogram (jagged thin curve) and true SDF (smooth thick) for a time series of length $N = 2048$ from an AR(24) process
- periodogram and true SDF are plotted on a decibel (dB) scale; i.e., $10 \log_{10} S(f)$ is plotted versus f
- bias (due to a phenomenon usually called ‘leakage’) is evident in the periodogram at high frequencies, where it differs from the true SDF by as much as 40 dB (i.e., four orders of magnitude!)

SDF Estimation via Periodogram: III

- can formulate SDF estimation as a 'signal + noise' problem
- $\hat{S}^{(p)}(f)$ itself is a signal $S(f) \times \chi_2^2$ noise
- usually $S(f) > 0$ and $\chi_2^2 > 0$, so can use a log transform to convert multiplicative model to additive model
- distribution of $\log \hat{S}^{(p)}(f)$ is the same as that of

$$\log\left(S(f)\chi_2^2/2\right) = \log\left(S(f)\right) + \log\left(\chi_2^2/2\right)$$

- Bartlett & Kendall (1946) show that

$$E\left\{\log\left(\chi_2^2/2\right)\right\} = -\gamma \text{ and } \operatorname{var}\left\{\log\left(\chi_2^2/2\right)\right\} = \pi^2/6$$

($\gamma \doteq 0.57721$ is Euler's constant), yielding

$$E\{\log(\hat{S}^{(p)}(f))\} = \log(S(f)) - \gamma \;\; \& \;\; \operatorname{var}\{\log(\hat{S}^{(p)}(f))\} = \pi^2/6$$

SDF Estimation via Periodogram: IV

- for $f_j = j/N$, model $Y^{(p)}(f_j) \equiv \log \hat{S}^{(p)}(f_j) + \gamma$ as

$$Y^{(p)}(f_j) = \log S(f_j) + \epsilon(f_j), \quad 0 < f_j < 1/2$$

  – regard $Y^{(p)}(f_j)$ as observed 'time' series
  – regard $\log S(f_j)$ as unknown signal
  – regard $\epsilon(f_j)$ as noise
    * $E\{\epsilon(f_j)\} = 0$ and $\text{var}\,\{\epsilon(f_j)\} = \pi^2/6$ (known!)
    * if $\{X_t\}$ is Gaussian, uncorrelatedness of $\hat{S}^{(p)}(f_j)$'s says that $\epsilon(f_j)$'s are uncorrelated
    * distribution of $\epsilon(f_j)$ is $\log(\chi_2^2)$ (markedly non-Gaussian)
- now have 'signal + noise' problem fitting form $\mathbf{Y} = \mathbf{D} + \boldsymbol{\epsilon}$

SDF Estimation via Periodogram: V

- Gao (1993) and Moulin (1994): estimate log SDF based upon $\mathcal{W}\mathbf{Y} = \mathcal{W}\mathbf{D} + \mathcal{W}\boldsymbol{\epsilon} \equiv \mathbf{d} + \mathbf{e}$
- $\boldsymbol{\epsilon}$ is IID, but non-Gaussian & hence same true of $\mathbf{e}$
- cannot use Gaussian-based universal threshold $\delta^{(u)}$
- basic steps in estimation procedure are the following
- assume $N = 2^J$ and use FFT algorithm to compute

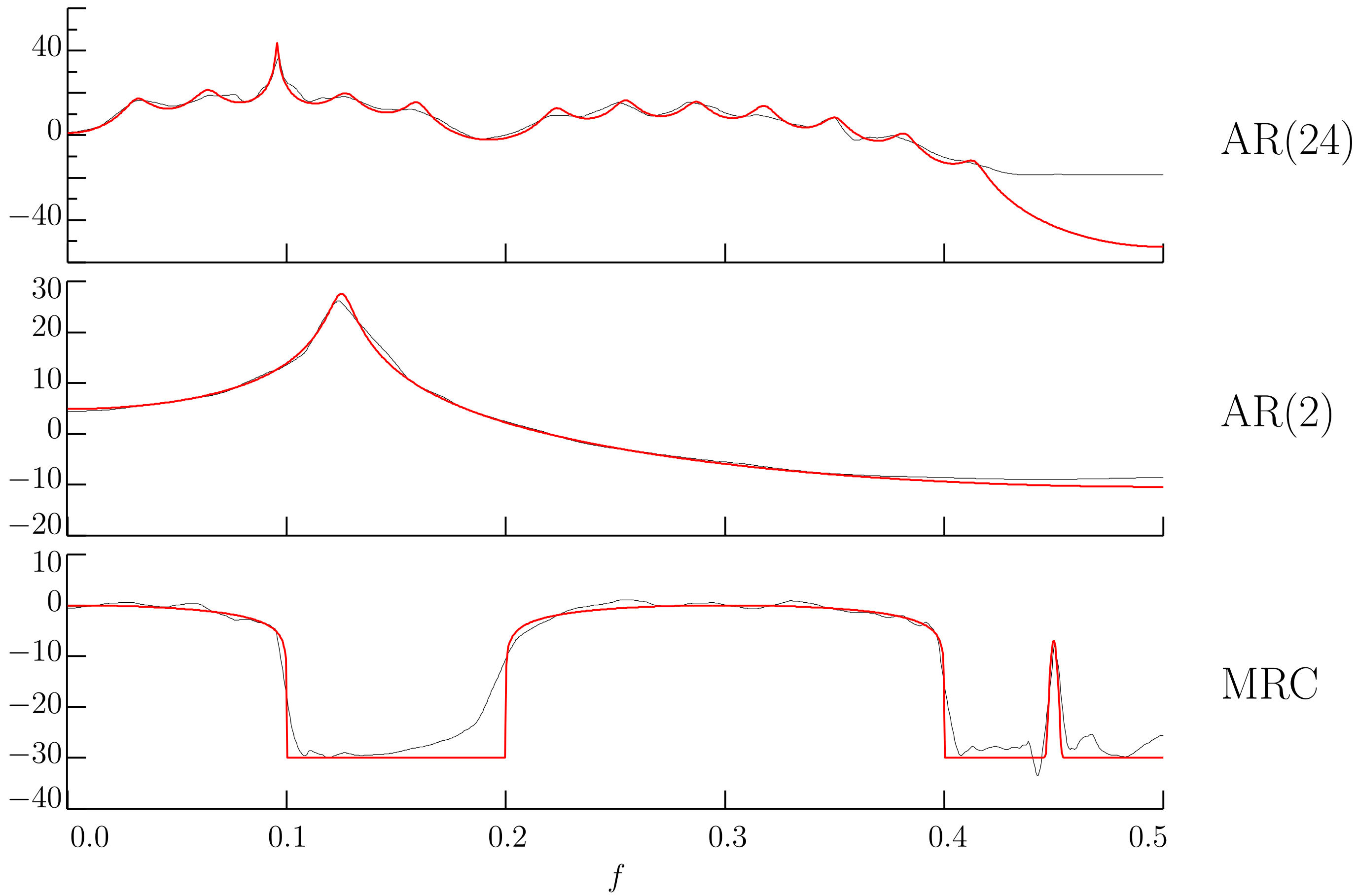
$$\hat{Y}^{(p)}(f_j) = \log \hat{S}^{(p)}(f_j) + \gamma, \quad f_j = \frac{j}{N},\ 0 \leq j \leq \frac{N}{2} - 1;$$

use of $\hat{Y}^{(p)}(f_0)$ not strictly OK, but small effect for large N

SDF Estimation via Periodogram: VI

- compute level J_0 partial DWT to obtain coefficients
$$\mathbf{W}_1^{(p)}, \mathbf{W}_2^{(p)}, \ldots, \mathbf{W}_{J_0}^{(p)} \text{ and } \mathbf{V}_{J_0}^{(p)},$$
where $\mathbf{W}_j^{(p)}$ has elements $W_{j,t}^{(p)} = d_{j,t} + e_{j,t}$
- apply thresholding scheme to $W_{j,t}^{(p)}$ to get $W_{j,t}^{(t)}$
  - for large j, can use (via 'central limit theorem' argument)
$$\delta^{(u)} = \left(2\sigma_\epsilon^2 \log\left(\frac{N}{2}\right)\right)^{1/2} = \left(2\frac{\pi^2}{6} \log\left(\frac{N}{2}\right)\right)^{1/2};$$
  - for small j, complicated methods required
- estimate $Y(f_j)$ by inverse transforming
$$\mathbf{W}_1^{(t)}, \mathbf{W}_2^{(t)}, \ldots, \mathbf{W}_{J_0}^{(t)} \text{ and } \mathbf{V}_{J_0}^{(p)}$$

Examples of SDF Estimation via Periodogram



- SDF estimates (thin jagged) and true SDFs (thick smooth) for AR(24), AR(2) and mobile radio communications processes

SDF Estimation via Multitapering: I

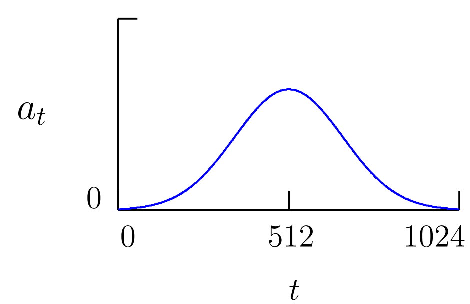
- refinement: use multitaper spectral estimator
- advantages of multitaper approach:
  - less biased than periodogram
  - log of multitaper estimator closer to Gaussian
- disadvantage: errors term now correlated, but this correlation structure obeys a simple model
- multitapering (1980s) builds upon older idea of tapering (1950s)
- rationale for tapering is to correct for bias in periodogram due to leakage (recall AR(24) periodogram)

SDF Estimation via Multitapering: II

- idea is to multiply time series by a data taper $\{a_t\}$ and then essentially form the periodogram for the tapered series:

$$\hat{S}^{(d)}(f) \equiv \left| \sum_{t=0}^{N-1} a_t X_t e^{-i2\pi f t} \right|^2$$

- resulting estimator $\hat{S}^{(d)}(\cdot)$ is called a direct spectral estimator
- $\{a_t\}$ is typically a bell-shaped curve



SDF Estimation via Multitapering: III

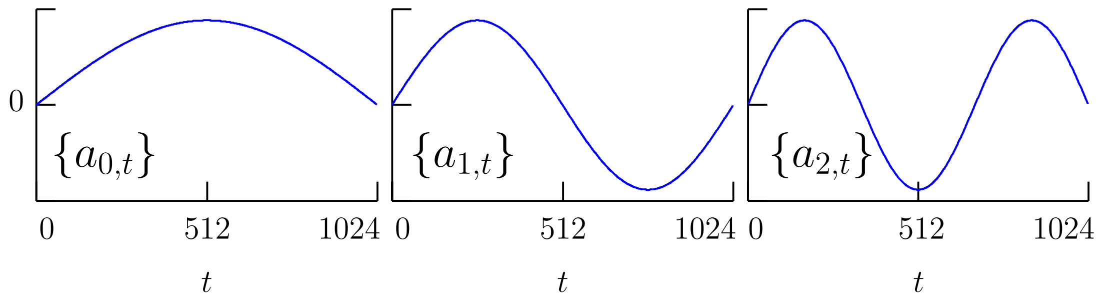
- critique of tapering is that it loses 'information' at end of series because sample size N is effectively shortened
- Thomson (1982): multitapering recovers 'lost info'
- idea is to use a set of K orthonormal data tapers $\{a_{n,t}\}$:

$$\sum_{t=0}^{N-1} a_{n,t} a_{l,t} = \begin{cases} 1, & \text{if } n = l; \\ 0, & \text{if } n \neq l. \end{cases} \quad 0 \leq n, l \leq K - 1$$

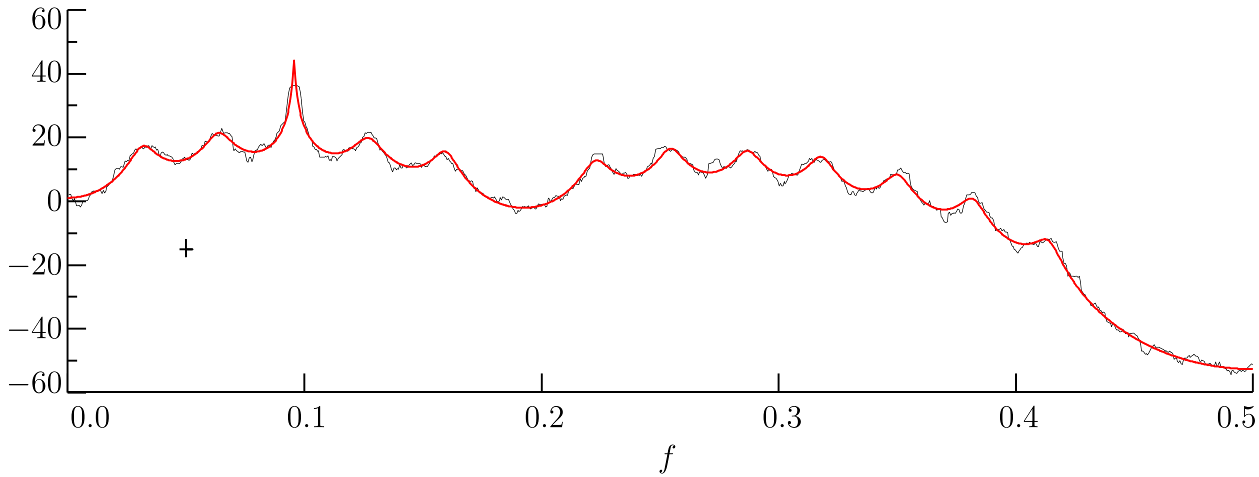
SDF Estimation via Multitapering: IV

- sine tapers are one possible set (Riedel & Sidorenko, 1995):

$$a_{n,t} = \left\{ \frac{2}{(N+1)} \right\}^{1/2} \sin \left\{ \frac{(n+1)\pi(t+1)}{N+1} \right\}, \quad t = 0, \ldots, N-1$$



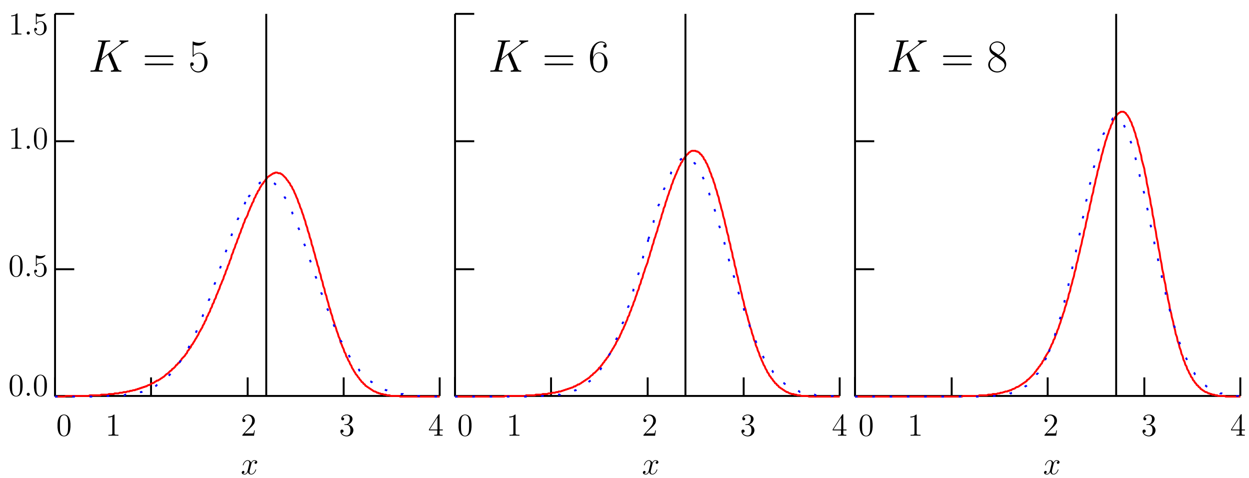
Example of SDF and Multitaper Estimator: I



- multitaper SDF estimate (thin jagged curve) and true SDF (thick smooth) for AR(24) time series of length $N = 2048$
- estimator based upon $K = 10$ sine tapers
- for large N and $0 < f < 1/2$, statistical theory says that $\hat{S}^{(mt)}(f)$ has a distribution given by $S(f)\chi^2_{2K}/2K$, where χ^2_{2K} is a chi-square RV with $2K$ degrees of freedom (DOFs)

Example of SDF and Multitaper Estimator: II

- for $K \geq 5$, distribution of $\log(\chi^2_{2K})$ is approximately Gaussian with mean $\psi(K) - \log(K)$ and variance $\psi'(K)$, where $\psi(\cdot)$ and $\psi'(\cdot)$ are the di- and trigamma functions
- solid curves are $\log(\chi^2_{2K})$ PDFs, while dotted curves are best approximating Gaussian PDFs

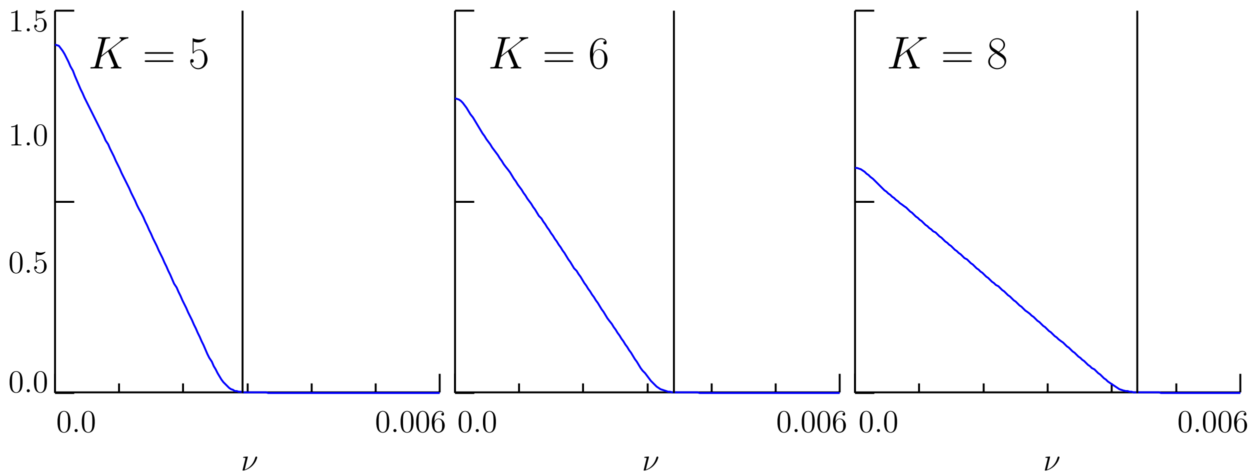


SDF Estimation via Multitapering: V

- model $Y^{(mt)}(f_j) \equiv \log \hat{S}^{(mt)}(f_j) - \psi(K) + \log(K)$ as
$$Y^{(mt)}(f_j) = \log S(f_j) + \eta(f_j), \;\; 0 < f_j < 1/2,$$
where now $f_j = j/2M$ with $2M \geq N$ (i.e., spacing of frequencies can be finer than that dictated by sample size N)
- similar to periodogram formulation of 'signal + noise' problem, but now fits the form $\mathbf{Y} = \mathbf{D} + \boldsymbol{\eta}$, where $\boldsymbol{\eta}$ is approximately zero mean Gaussian (if $K \geq 5$), but correlated
- can argue that $\text{cov}\{\eta(f_j), \eta(f_k)\} \equiv s_\eta(f_j - f_k)$, i.e., depends on just 'lag' $\nu = f_j - f_k$

SDF Estimation via Multitapering: VI

- $s_\eta(\nu)$ is approximately 'triangular', with a cutoff dictated by the bandwidth $\frac{K+1}{N+1}$ associated with the multitaper estimator



SDF Estimation via Multitapering: VII

- covariance matrix $\Sigma_{\boldsymbol{\eta}}$ for $\boldsymbol{\eta}$ well approximated by the following 'circular' matrix dictated by $s_\eta(\cdot)$:

$$\begin{bmatrix} s_\eta(f_0) & \cdots & s_\eta(f_{\frac{M}{2}-1}) & s_\eta(f_{\frac{M}{2}}) & s_\eta(f_{\frac{M}{2}-1}) & \cdots & s_\eta(f_1) \\ s_\eta(f_1) & \cdots & s_\eta(f_{\frac{M}{2}-2}) & s_\eta(f_{\frac{M}{2}-1}) & s_\eta(f_{\frac{M}{2}}) & \cdots & s_\eta(f_2) \\ s_\eta(f_2) & \cdots & s_\eta(f_{\frac{M}{2}-3}) & s_\eta(f_{\frac{M}{2}-2}) & s_\eta(f_{\frac{M}{2}-1}) & \cdots & s_\eta(f_3) \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ s_\eta(f_1) & \cdots & s_\eta(f_{\frac{M}{2}}) & s_\eta(f_{\frac{M}{2}-1}) & s_\eta(f_{\frac{M}{2}-2}) & \cdots & s_\eta(f_0) \end{bmatrix}$$

- leads to following procedure for estimating $S(\cdot)$

SDF Estimation via Multitapering: VIII

- let $M \geq N/2$ be any power of 2, i.e., $M = 2^q$
- compute $\hat{S}^{(mt)}(\cdot)$ on tapered series padded with $2M - N$ zeros

$$\{a_{k,0}X_0, \ldots, a_{k,N-1}X_{N-1}, 0, \ldots, 0\}, \quad k = 0, \ldots, K-1$$

- form $Y^{(mt)}(f_j) \equiv \log \hat{S}^{(mt)}(f_j) - \psi(K) + \log(K)$ with $f_j = j/2M$
- compute level J_0 partial DWT for $Y^{(mt)}(f_j)$, $0 \leq f_j < 1/2$:

$$\mathbf{W}_1^{(mt)}, \mathbf{W}_2^{(mt)}, \ldots, \mathbf{W}_{J_0}^{(mt)} \text{ and } \mathbf{V}_{J_0}^{(mt)}$$

elements of $\mathbf{W}_j^{(mt)}$ are $W_j^{(mt)} = d_{j,t} + n_{j,t}$

- can show that $\text{var}\{n_{j,t}\} \equiv \frac{1}{M}\sum_{k=0}^{M-1} S_k \mathcal{H}_j(\frac{k}{M}) \equiv \sigma_j^2$, where $\{S_k\}$ is DFT of first row of circular approximation to $\Sigma_{\boldsymbol{\eta}}$, and $\mathcal{H}_j(\cdot)$ is squared gain for jth level equivalent filter $\{h_{j,l}\}$

SDF Estimation via Multitapering: IX

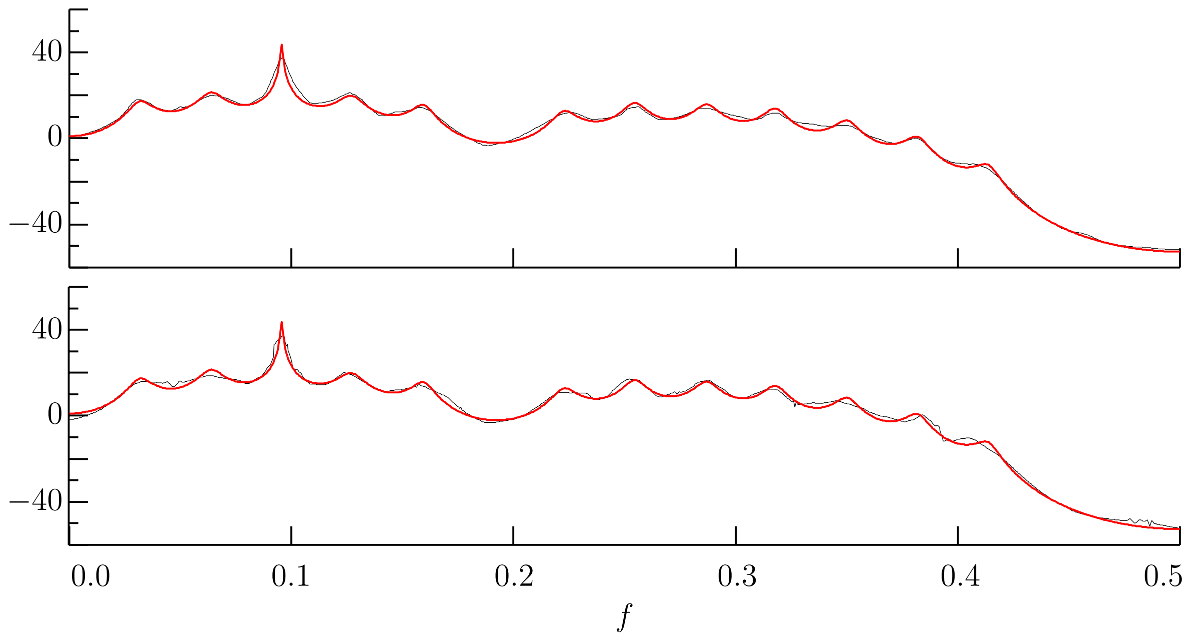
- can show that $\sigma_j^2 < \sigma_{j+1}^2$, i.e., variance increases with scale
- can show that $\sigma_p^2 < \sigma_\eta^2 = \psi'(K) \leq \sigma_{p+1}^2$ for some p; e.g., for Haar, $p = 2$ for $5 \leq K \leq 10$
- apply thresholding to $\mathbf{W}_j^{(mt)}$ to obtain $\mathbf{W}_j^{(t)}$ using either

  1. level/scale dependent thresholds $\delta_j = (2\sigma_j^2 \log \frac{N}{2})^{1/2}$ or
  2. level/scale independent thresholds $\delta = (2\psi'(K) \log \frac{N}{2})^{1/2}$

- 2nd scheme will suppress small scale 'noise spikes' while leaving 'informative' coarse scale coefficients relatively unattenuated
- estimate log SDF by inverse transforming

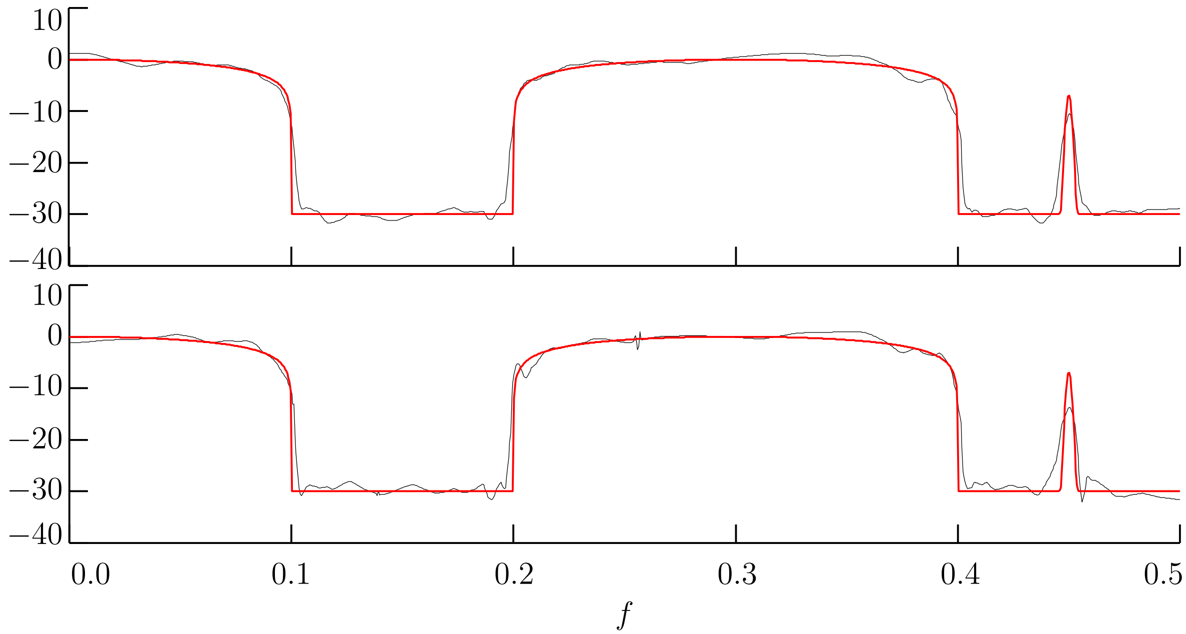
$$\mathbf{W}_1^{(t)}, \mathbf{W}_2^{(t)}, \ldots, \mathbf{W}_{J_0}^{(t)} \text{ and } \mathbf{V}_{J_0}^{(mt)}$$

Examples of Estimation via Multitapering: I



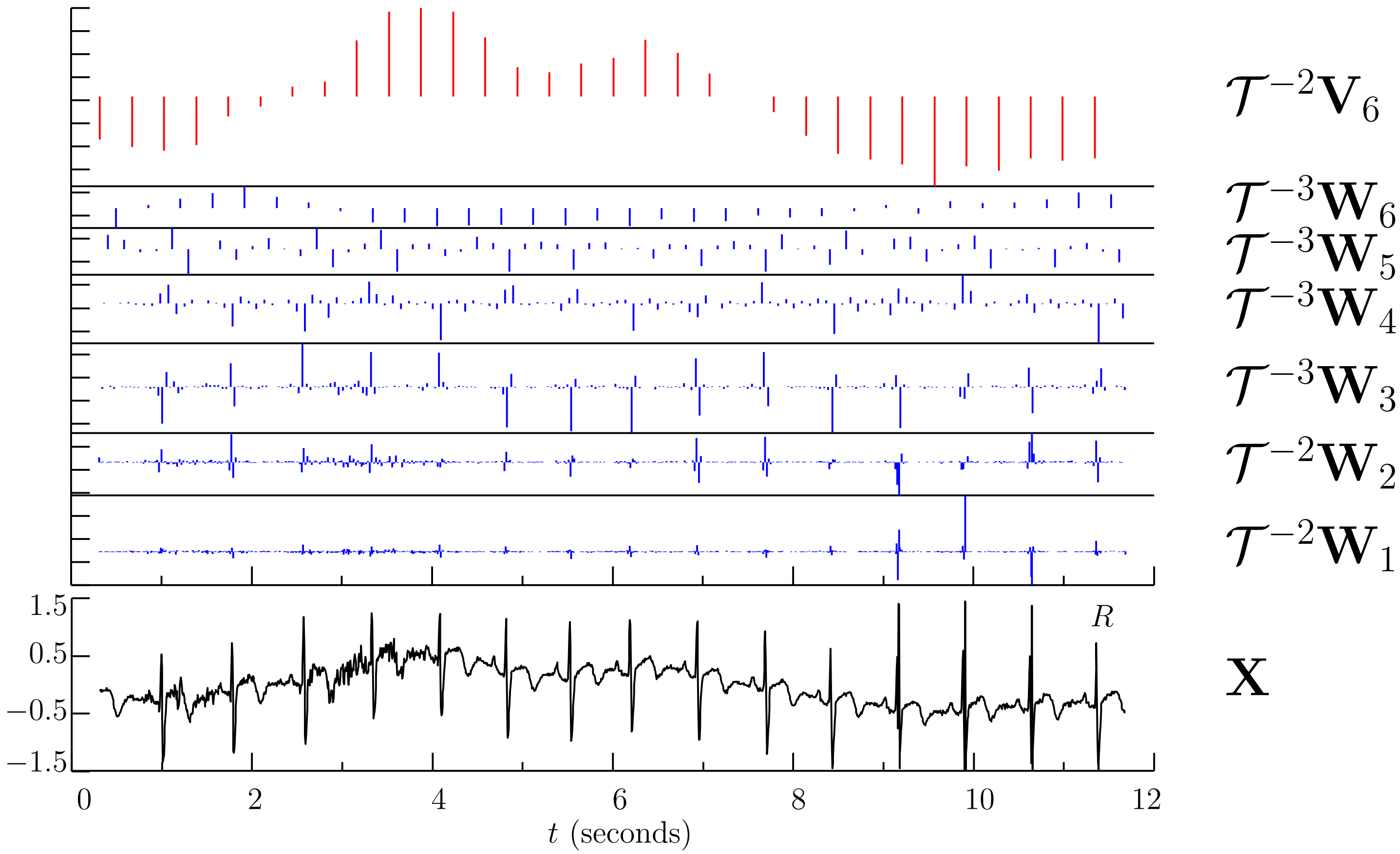
- estimated/true SDFs (thin jagged/thick smooth curves)
- estimates are 'representative' in having RMSEs closest to the average RMSE over 1000 simulations (each with $N = 2048$)
- upper: level-independent soft thresholding; lower: dependent & hard ($J_0 = 5$ LA(8) DWT with $K = 10$ sine multitapers)

Examples of Estimation via Multitapering: II



- as in previous figure, but now for the mobile radio communications process
- computer experiments show multitaper-based estimator outperforms periodogram scheme for AR(24), AR(2) and MRC processes considered by Gao (1993) and Moulin (1994)

Comments on 'Second Generation' Denoising: I



- 'classical' denoising looks at each $W_{j,t}$ alone; for 'real world' signals, coefficients often cluster within a given level and persist across adjacent levels (ECG series offers an example)

Comments on 'Second Generation' Denoising: II

- here are some 'second generation' approaches that exploit these 'real world' properties:
  - Crouse *et al.* (1998) use hidden Markov models for stochastic signal DWT coefficients to handle clustering, persistence and non-Gaussianity
  - Huang and Cressie (2000) consider scale-dependent multiscale graphical models to handle clustering and persistence
  - Cai and Silverman (2001) consider 'block' thesholding in which coefficients are thresholded in blocks rather than individually (handles clustering)
  - Dragotti and Vetterli (2003) introduce the notion of 'wavelet footprints' to track discontinuities in a signal across different scales (handles persistence)

Comments on 'Second Generation' Denoising: III

- 'classical' denoising also suffers from problem of overall significance of multiple hypothesis tests
- 'second generation' work integrates idea of 'false discovery rate' (Benjamini and Hochberg, 1995) into denoising (see Wink and Roerdink, 2004, for an applications-oriented discussion)
- for some second generation developments, see
  - review article by Antoniadis (2007)
  - Chapters 3 and 4 of book by Nason (2008)
  - October 2009 issue of *Statistica Sinica*, which has a special section entitled 'Multiscale Methods and Statistics: A Productive Marriage'

Additional References

- A. Antoniadis (2007), 'Wavelet Methods in Statistics: Some Recent Developments and Their Applications,' *Statistical Surveys*, **1**, pp. 16–55
- Y. Benjamini and Y. Hochberg (1995), 'Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing,' *Journal of the Royal Statistical Society, Series B*, **57**, pp. 289–300
- T. Cai and B. W. Silverman (2001), 'Incorporating Information on Neighboring Coefficients into Wavelet Estimation,' *Sankhya Series B*, **63**, pp. 127–48
- P. L. Dragotti and M. Vetterli (2003), 'Wavelet Footprints: Theory, Algorithms, and Applications,' *IEEE Transactions on Signal Processing*, **51**, pp. 1306–23
- H.–C. Huang and N. Cressie (2000), 'Deterministic/Stochastic Wavelet Decomposition for Recovery of Signal from Noisy Data,' *Technometrics*, **42**, pp. 262–76
- G. P. Nason (2008), *Wavelet Methods in Statistics with R*, Springer, Berlin
- A. M. Wink and J. B. T. M. Roerdink (2004), 'Denoising Functional MR Images: A Comparison of Wavelet Denoising and Gaussian Smoothing,' *IEEE Transactions on Medical Imaging*, **23**(3), pp. 374–87