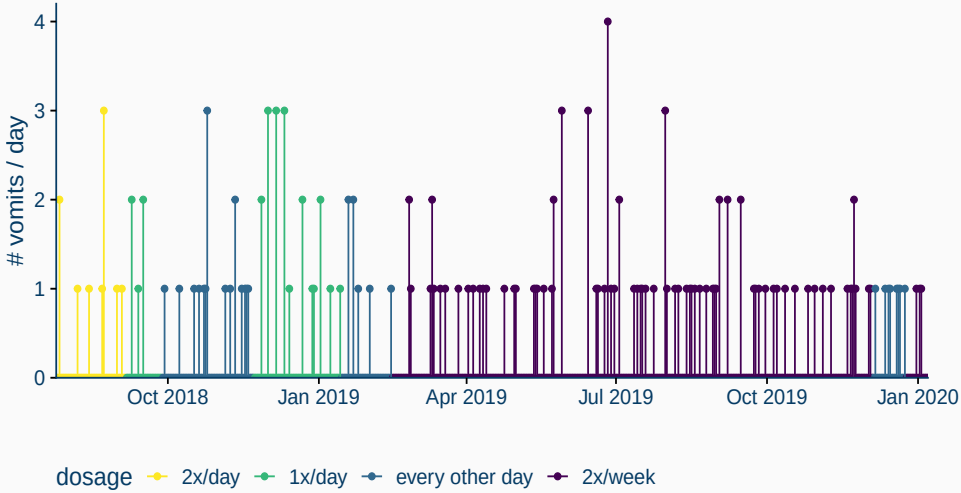
Trend filtering in exponential families

Daniel J. McDonald
Indiana University, Bloomington
dajmcdon.github.io

7 January 2020

These are my cats

Number of vomits/day



Poisson model

y_i is the number of vomits on day i

Poisson distributed with time-varying parameter ϕ_i

$L(\phi \mid y) = \prod_{i=1}^{n} \frac{\phi_i^{y_i} \exp(-\phi_i)}{y_i!}$

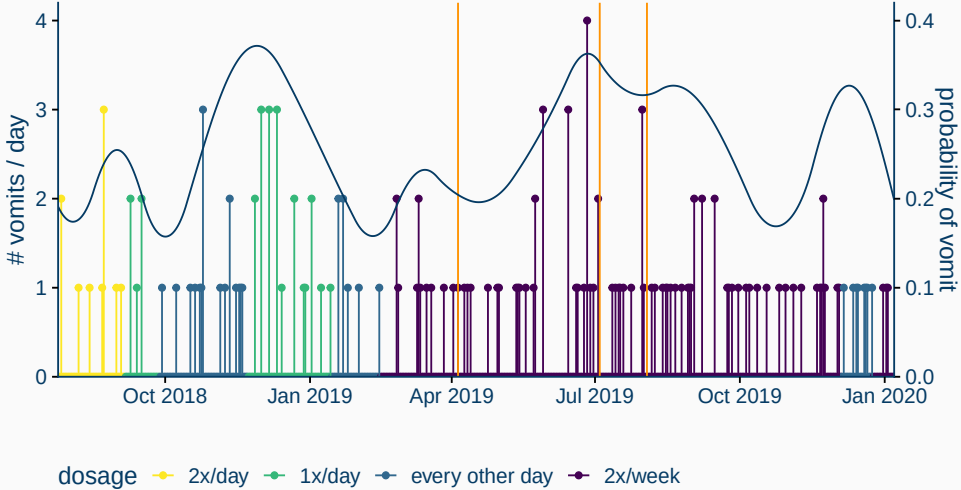
Goal: estimate ϕ from data, ϕ should be "smooth".

Set $\theta_i = \log \phi_i$

$$\underset{\theta}{\text{minimize}} \ \mathbf{1}^\top \exp(\theta_i) - y^\top \theta + \lambda \left\| D\theta \right\|_1$$

D matrix encodes smoothness

Trend filtering



What's this talk about?

Trend filtering is not new.

Aside from small specializations,

- the theory is for Gaussian mean
- the algorithms are for Gaussian mean on grids or tree-like graphs
- the implementations work on "small" data
- λ selection is for Gaussian mean

See Hütter and Rigollet (2016); Kim et al. (2009); Sadhanala et al. (2017); Tibshirani (2014); Wang et al. (2016)

What's this talk about?

We generalize to exponential families

1. Provide some algorithms that work on big data
2. Select λ reasonably
3. Near-minimax theoretical guarantees

What's this talk about?

We generalize to exponential families

1. Provide some algorithms that work on big data
2. Select λ reasonably
3. Near-minimax theoretical guarantees

Motivated by a climate change study

Estimating the trend in cloud-top temperature volatility

Climate change

The scientific consensus is that

1. World-wide climate is changing.
2. This change is mostly driven by human behavior.

~~Global warming~~ $\longrightarrow$ climate change: the distribution of temperature (and precipitation) is changing

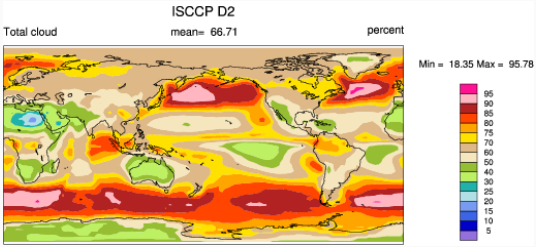
Increasing mean temperature understates the costs:

1. More frequent extremes have severe effects
2. Local discrepancies lead to more storms
3. Temporal dependencies imply persistence

Using weather satellites

Drivers of climate variation:

1. Ocean currents
2. Jet stream
3. Annular modes
4. Cloudiness

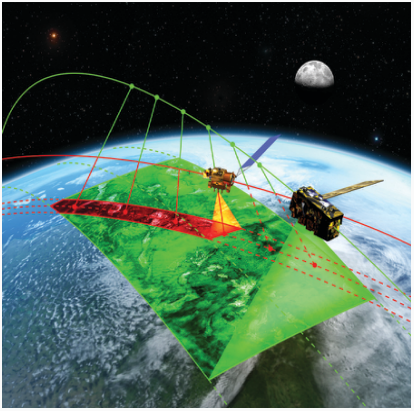


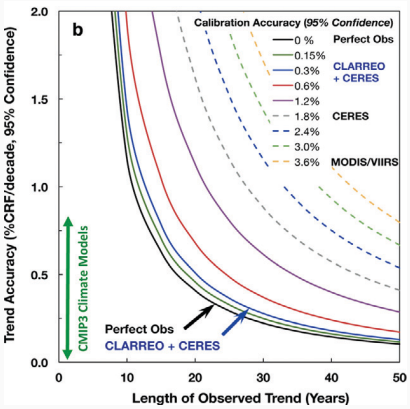
CLARREO satellite: monitor cloud top temperature as it relates to climate.

- Originally slated to launch in 2020
- Trump Administration killed it in 2017
- Revived by NASA last year
- Launching no sooner than 2023

Source: NCAR CCSM3 Diagnostic Plots.

CLARREO vs MetOp/Modis





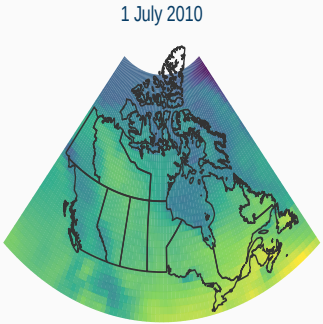
- Weather satellites aren't made for this.
- More information in higher moments than in average?

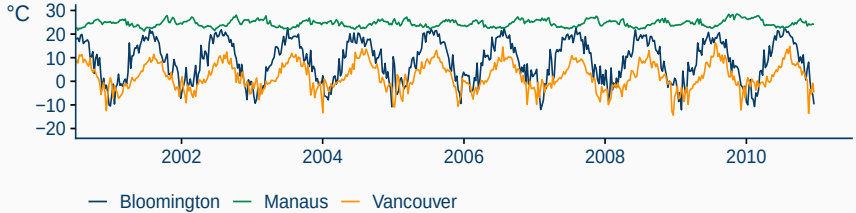
Source: Wielicki, et al. (2013).

Satellite data

Once collaborators do lots of processing…

- 52,000 time series
- daily records over ∼ 50 years
- “trends” are local, nonlinear, not sinusoidal



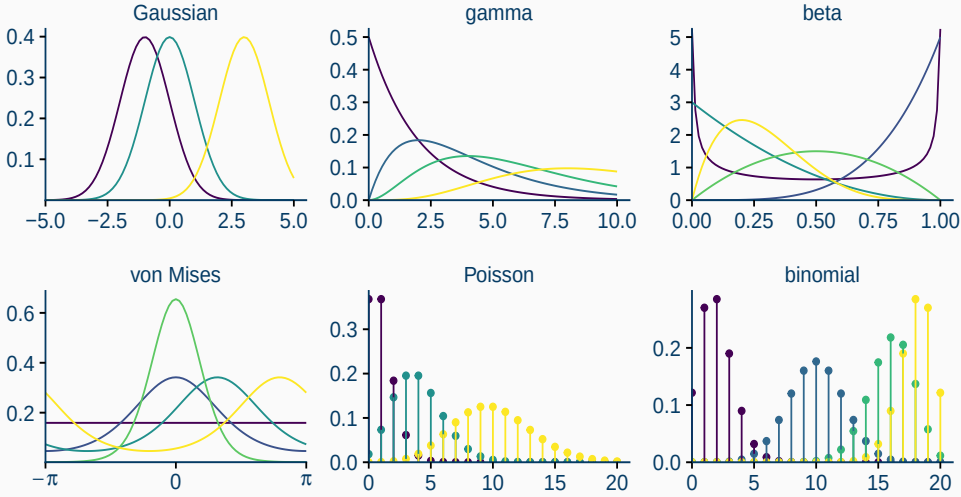


Trends in variance

- Let X_{ijt} be the observed temperature at time t and location (i, j).
- Suppose $X_{ijt} \sim \text{Normal}\left(0, \sigma^2_{ijt}\right)$
- (Follows sophisticated detrending)
- Estimate σ^2, but it should be "smooth" relative to space and time.
- Use a matrix $D+$ penalty to encode this smoothness.

Exponential families

Standard examples



Natural exponential family

Let X be a random variable with pdf/pmf $f_X(x; \phi)$

If I can write

$$f_X(x) = h(x) \exp\left(y(x) \cdot \theta(\phi) - A(\theta)\right)$$

Then, X belongs to the (single parameter) exponential family of distributions

Using (Y, θ) instead of (X, ϕ) is the "natural" parameterization

Natural exponential family

Let X be a random variable with pdf/pmf $f_X(x; \phi)$

If I can write

$$f_X(x) = h(x) \exp\Big(y(x) \cdot \theta(\phi) - A(\theta)\Big)$$

Then, X belongs to the (single parameter) exponential family of distributions

Using (Y, θ) instead of (X, ϕ) is the "natural" parameterization

Some nice properties

1. A is convex for all $\theta_0 \in \Theta := \left\{\theta : \int_y \exp(y\theta) d\mu(y) < \infty\right\}$
2. All derivatives of A exist at $\theta_0 \in \text{int}(\Theta)$
3. $\mathbb{E}[Y] = A'(\theta_0), \quad \mathbb{V}[Y] = A''(\theta_0)$
4. $KL(\theta_0 \,||\, \theta_1) = A(\theta_1) - A(\theta_0) - A'(\theta_0)^\top(\theta_1 - \theta_0)$

Trend filtering

Optimization problem

General: $Y_i \sim \text{ExpFam}(\theta_i)$

$$\min_{\theta \in \Theta} \; \mathbf{1}^\top A(\theta) - y^\top \theta + \lambda \left\| D\theta \right\|_1$$

Optimization problem

General: $Y_i \sim \text{ExpFam}(\theta_i)$

$$\min_{\theta \in \Theta} \ \mathbf{1}^\top A(\theta) - y^\top \theta + \lambda \left\| D\theta \right\|_1$$

Gaussian: $X_i \sim \text{N}(\mu_i,\ 1)$

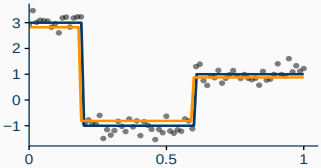
$$\min_{\mu \in \mathbb{R}^n} \ \frac{1}{2} \left\| x - \mu \right\|_2^2 + \lambda \left\| D\mu \right\|_1 = \min_{\theta \in \mathbb{R}^n} \ \frac{1}{2} \theta^\top \theta - y^\top \theta + \lambda \left\| D\theta \right\|_1$$

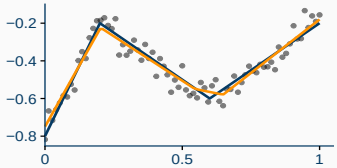
Gaussian: $X_i \sim \text{N}(0,\ \sigma_i^2)$

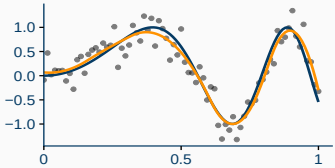
$$\min_{\theta \in (-\infty, 0)^n} \ -\frac{1}{2} \mathbf{1}^\top \log(-\theta) - y^\top \theta + \lambda \left\| D\theta \right\|_1$$

$\theta = -\frac{1}{2\sigma^2}$, $y = x^2$, and $A(z) = -\frac{1}{2} \log(-z)$

Smoothness and penalty order, *D* matrices







$$\begin{bmatrix} -1 & 1 & & & & 0 \\ & -1 & 1 & & & \\ & & \ddots & \ddots & & \\ & & & \ddots & \ddots & \\ 0 & & & & -1 & 1 \\ & & & & & -1 & 1 \end{bmatrix}$$

Constant, k=0

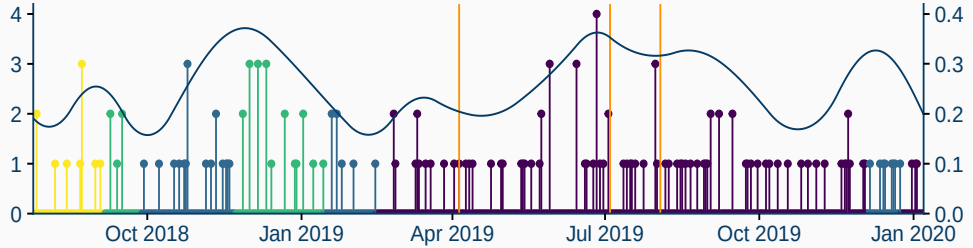
$$\begin{bmatrix} 1 & -2 & 1 & & & 0 \\ & 1 & -2 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & \ddots & \ddots & \ddots \\ 0 & & & & 1 & -2 & 1 \\ & & & & & 1 & -2 & 1 \end{bmatrix}$$

Linear, k=1

$$\begin{bmatrix} -1 & 3 & -3 & 1 & & & 0 \\ & -1 & 3 & -3 & 1 & & \\ & & \ddots & \ddots & \ddots & \ddots & \\ & & & \ddots & \ddots & \ddots & \ddots \\ 0 & & & & -1 & 3 & -3 & 1 \\ & & & & & -1 & 3 & -3 & 1 \end{bmatrix}$$

Quadratic, k=2

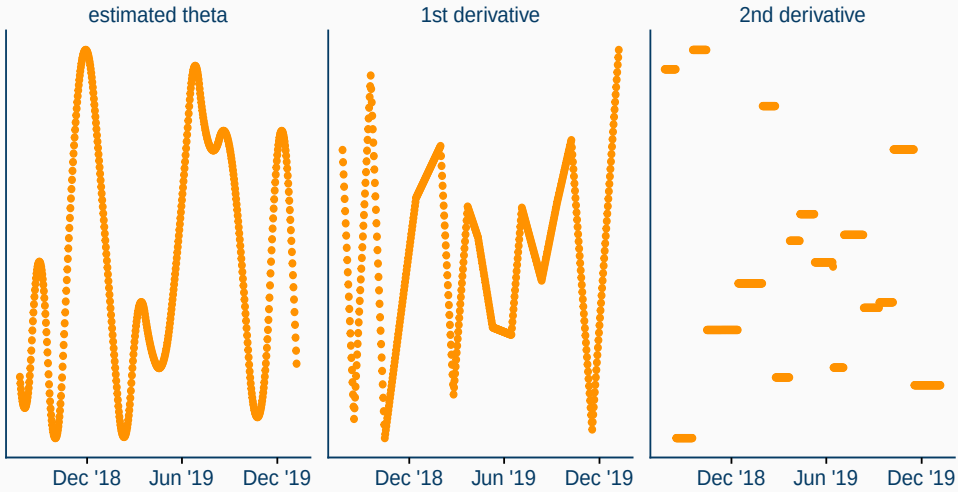
Quadratic Poisson trend filtering



Looks visually like a smoothing spline, but more locally adaptive

Works well on functions of "bounded variation": $\int_{\mathcal{X}} |\theta^{(k)}(x)| dx < \infty$

Derivative properties



Relations to other (similar) methods

Locally adaptive regression splines

$$\min_{f \in \mathcal{F}_k} \frac{1}{2n} \|y - f\|_2^2 + \lambda \mathrm{TV}(f^{(k)})$$

- $k = 0,\ 1$ is equivalent to TF; $k \geq 2$, equivalent as $n \longrightarrow \infty$
- TF computations cost $O(n)$ compared to $O(n^3)$

Smoothing splines

$$\min_{f \in \mathcal{W}_{(k+1)/2}} \frac{1}{2n} \|y - f\|_2^2 + \lambda \int_{\mathcal{X}} \left(f^{\left(\frac{k+1}{2}\right)}(t) \right)^2 dt$$

- Similar computational burden (if B-spline basis)
- TF is more adaptive for equivalent complexity

(Green and Silverman, 1994; Mammen and van de Geer, 1997; Wahba, 1990)

Complexity

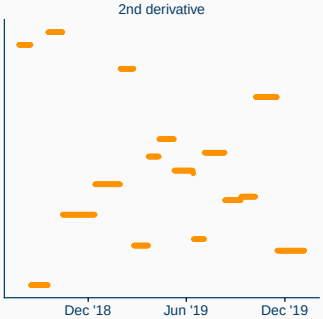
The Degrees of Freedom measures “complexity”

Think OLS: p predictors and intercept $\longrightarrow$ df $= p + 1$

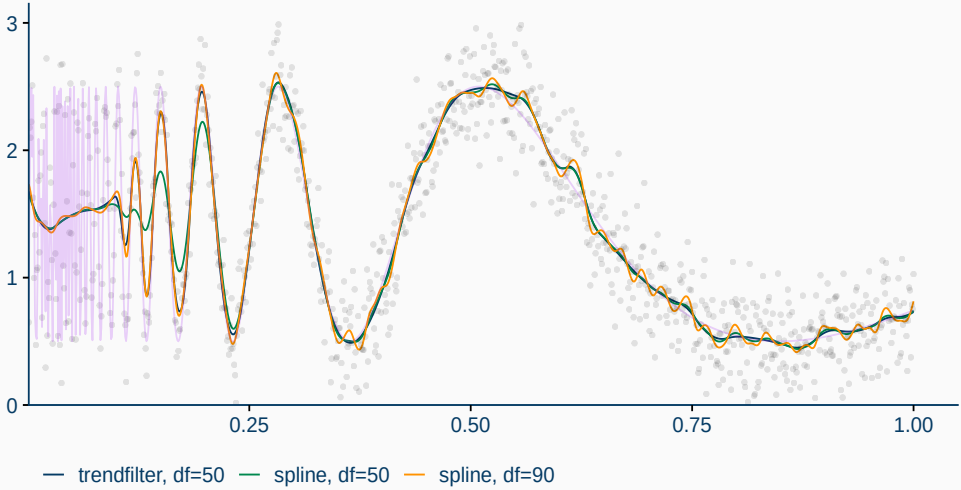
TF and Gaussian mean df $= \mathbb{E}\left[\#\text{ knots}\right] + k + 1$

$\widehat{df} = \#\text{ knots} + k + 1$

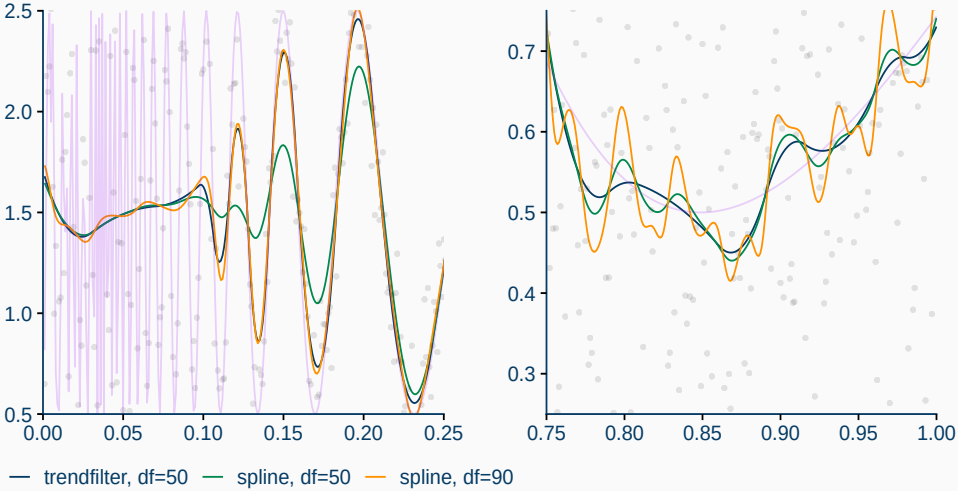
Smoothing splines have same degrees of freedom



Local adaptivity



Local adaptivity



Algorithms

Optimization problem

$$\min_{\theta} \mathbf{1}^\top A(\theta) - y^\top \theta + \lambda \left\| D\theta \right\|_1$$

Standard optimizer: Primal Dual Interior Point method

Alternatively (for $k \geq 1$): Alternating Direction Method of Multipliers

see Kim et al. (2009); Tibshirani (2014)

Generic Primal Dual Interior Point

1. Start with a guess $\theta^{(1)}$
2. Solve a linear system $[Ms = v]$
3. Calculate a step size
4. Iterate 2 & 3 until convergence

Generic Primal Dual Interior Point

1. Start with a guess $\theta^{(1)}$
2. Solve a linear system $[Ms = v]$
3. Calculate a step size
4. Iterate 2 & 3 until convergence

M is a function of D and θ

Banded for TF

So 2 and 3 are solved in linear time.

Alternating direction method of multipliers

Restate the problem

Original	Equivalent
$\min_x \quad f(x) + g(x)$	$\min_{x,z} \quad f(x) + g(z)$ s.t. $\quad x - z = 0$

Then, iterate the following:

$$x \leftarrow \operatorname*{argmin}_x f(x) + \frac{\rho}{2} \|x - z + u\|_2^2$$

$$z \leftarrow \operatorname*{argmin}_z g(z) + \frac{\rho}{2} \|x - z + u\|_2^2$$

$$u \leftarrow u + x - z$$

Why would you do this?

Decouples f and g

If f and g are nice, can be parallelized

Converges under very general conditions

Often many ways to decouple a problem

Decoupling example (Gaussian mean)

Original

$$\min_{\theta} \quad \frac{1}{2}\theta^\top\theta - y^\top\theta + \lambda \left\|D\theta\right\|_1$$

Equivalent

$$\min_{\theta,\alpha} \quad \frac{1}{2}\theta^\top\theta - y^\top\theta + \lambda \left\|\alpha\right\|_1$$

$$\text{s.t.} \quad D\theta - \alpha = 0$$

$$\theta \leftarrow \operatorname*{argmin}_{\theta} \frac{1}{2}\theta^\top\theta - y^\top\theta + \frac{\rho}{2}\left\|\alpha - D\theta + u\right\|_2^2$$

$$\alpha \leftarrow \operatorname*{argmin}_{\alpha} \lambda \left\|\alpha\right\|_1 + \frac{\rho}{2}\left\|D\theta - \alpha + u\right\|_2^2$$

$$u \leftarrow u - D\theta + \alpha$$

Decoupling example (Gaussian mean)

Original

$$\min_{\theta} \quad \frac{1}{2}\theta^\top\theta - y^\top\theta + \lambda \left\|D\theta\right\|_1$$

Equivalent

$$\min_{\theta,\alpha} \quad \frac{1}{2}\theta^\top\theta - y^\top\theta + \lambda \left\|\alpha\right\|_1$$

$$\text{s.t.} \quad D\theta - \alpha = 0$$

$\theta \longleftarrow$ matrix multiply

$\alpha \longleftarrow$ elementwise soft-threshold

$u \longleftarrow$ add vectors

Decoupling example (Gaussian mean)

Original

$$\min_{\theta} \quad \frac{1}{2}\theta^\top\theta - y^\top\theta + \lambda \left\|D\theta\right\|_1$$

Equivalent

$$\min_{\theta,\alpha} \quad \frac{1}{2}\theta^\top\theta - y^\top\theta + \lambda \left\|\alpha\right\|_1$$

$$\text{s.t.} \quad D\theta - \alpha = 0$$

$$\theta \leftarrow \left(I_n + \rho D^\top D\right)^{-1}\left(y + \rho D^\top(\alpha + u)\right)$$

$$\alpha \leftarrow \mathcal{S}_{\lambda/\rho}(D\theta + u)$$

$$u \leftarrow u - D\theta + \alpha$$

$[\mathcal{S}_a(b)]_k = \text{sign}(b_k)(|b_k| - a)_+$

What about for climate data?

Existing implementations of PDIP/ADMM are fast because D is banded, loss is quadratic

Climate data is over a 3D grid (lat $\times$ lon $\times$ time)

But not quite a grid because observations are on a sphere

So D is not banded and loss isn't quadratic

What about for climate data?

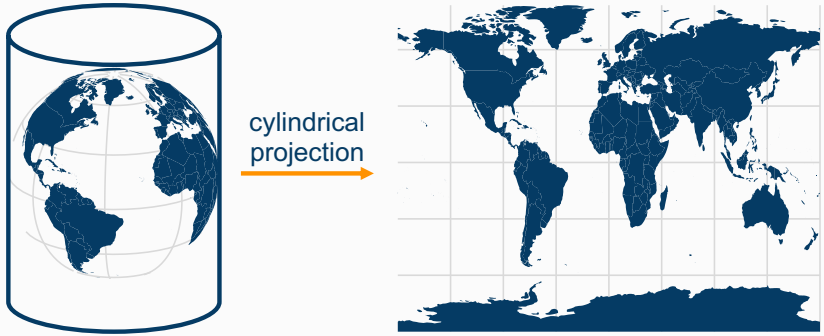
M is now dense and $10^9 \times 10^9$

M occupies 8000 Petabytes, and you have to invert it

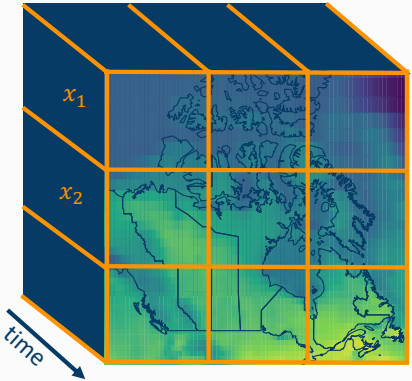
Same issue for existing ADMM implementations

Need custom algorithms/code

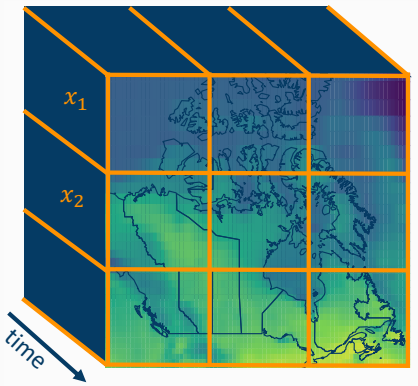
What our data look like



Consensus version



Consensus version



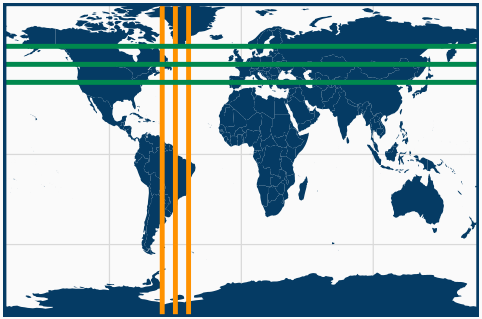
$x_g \longleftarrow$ use PDIP on smaller blocks

$\theta \longleftarrow$ average over groups

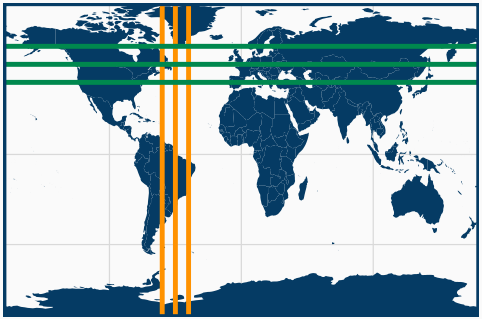
$u_g \longleftarrow$ add vectors

Requires very few iterations, but iterations cost $O\left(|\text{block}|^3\right)$. Can parallelize over blocks.

Grid world



Grid world



$\theta_{ijt} \longleftarrow$ find a root

each line $\longleftarrow$ 1D TF with the convex loss

dual variables $\longleftarrow$ add vectors

Requires many iterations, but iterations cost $O(|\text{line}|)$. Can parallelize over lines.

Other algorithmic issues

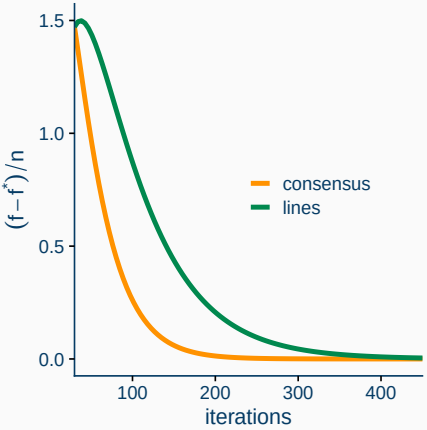
Other possible versions depending on architecture

Off-the-shelf stuff doesn't work

Simulations: 4 sec vs 2 hours at 400 iterations

Smaller problems don't need these details

Must repeat for many tuning parameters



(Khodadadi and McDonald, 2019)

Tuning parameter selection

Minimize risk

- Suppose you have:
  - observed data $y \sim (\theta_o, \sigma^2 I_n)$
  - an estimator $\widehat{\theta}_\lambda$
- You want $\text{MSE}(\lambda) = \mathbb{E}_Y \left[\left\| \theta_o - \widehat{\theta}_\lambda \right\|_2^2 \right]$
- $\text{Error}(\lambda) = \left\| y - \widehat{\theta}_\lambda \right\|_2^2$ is biased
- AIC, BIC, GCV compensate with $\text{Error}(\lambda) + \text{pen}(\lambda)$
- Cross Validation uses held-out sets

Unbiased estimation

If $Y \sim (\theta_0, \sigma^2 I_n)$, then

$$\begin{aligned} \text{MSE}(\lambda) &= \mathbb{E}_Y \left[\left\| \theta_0 - \widehat{\theta}_\lambda(Y) \right\|_2^2 \right] \\ &= \mathbb{E} \left[\left\| Y - \widehat{\theta}_\lambda(Y) \right\|_2^2 \right] - n\sigma^2 + 2\text{tr Cov}(Y,\ \widehat{\theta}_\lambda(Y)) \end{aligned}$$

e.g. Efron (1986)

Unbiased estimation

If $Y \sim (\theta_0, \sigma^2 I_n)$, then

$$\begin{aligned} \mathrm{MSE}(\lambda) &= \mathbb{E}_Y \left[\left\| \theta_0 - \widehat{\theta}_\lambda(Y) \right\|_2^2 \right] \\ &= \mathbb{E} \left[\left\| Y - \widehat{\theta}_\lambda(Y) \right\|_2^2 \right] - n\sigma^2 + 2\mathrm{tr}\ \mathrm{Cov}(Y,\ \widehat{\theta}_\lambda(Y)) \end{aligned}$$

Example:

$$\begin{aligned} \widehat{\theta}_\lambda(y) &= Wy \\ \mathrm{tr}\ \mathrm{Cov}\left(Y,\ \widehat{\theta}_\lambda(Y)\right) &= \sigma^2 \mathrm{tr}\,(W) \\ \widehat{\mathrm{MSE}}(\lambda) &= \left\| Y - \widehat{\theta}_\lambda(Y) \right\|_2^2 - n\sigma^2 + 2\mathrm{df} \end{aligned}$$

e.g. Efron (1986)

Stein's unbiased risk estimator

1. $Y \sim \text{Normal}\left(\theta_0,\ \sigma^2 I_n\right)$
2. $\widehat{\theta}_\lambda(\cdot)$ weakly differentiable with ess. bounded partials

$$\text{tr Cov}\left(Y,\ \widehat{\theta}_\lambda(Y)\right) = \sigma^2 \sum_i \mathbb{E}\left[\frac{\partial \widehat{\theta}_{\lambda,i}}{\partial Y_i}(Y)\right]$$

(Stein, 1981)

Stein's unbiased risk estimator

1. $Y \sim \text{Normal}\left(\theta_0,\ \sigma^2 I_n\right)$
2. $\widehat{\theta}_\lambda(\cdot)$ weakly differentiable with ess. bounded partials

$$\text{tr Cov}\left(Y,\ \widehat{\theta}_\lambda(Y)\right) = \sigma^2 \sum_i \mathbb{E}\left[\frac{\partial \widehat{\theta}_{\lambda,i}}{\partial Y_i}(Y)\right]$$

- Ingredients for Stein's Unbiased Risk Estimator:
  1. Expression for risk I want, w/o dependence on parameters
  2. Expression for $\mathbb{E}\left[\frac{\partial \widehat{\theta}_{\lambda,\,i}}{\partial Y_i}(Y)\right]$

(Stein, 1981)

Generalized SURE for continuous exp fam

1. $p_\theta(y) = h(y)\exp(\theta^\top y - \mathbf{1}^\top A(\theta))$
2. $h(\cdot)$ is weakly differentiable

$$\mathbb{E}\left[\theta_0^\top \widehat{\theta}_\lambda(Y)\right] = -\mathbb{E}\left[\left\langle \frac{h'(Y)}{h(Y)},\ \widehat{\theta}_\lambda(Y)\right\rangle + \sum_i \left(\frac{\partial \widehat{\theta}_{\lambda,i}}{\partial Y_i}(Y)\right)\right]$$

(Eldar, 2009)

Generalized SURE for continuous exp fam

1. $p_\theta(y) = h(y)\exp(\theta^\top y - \mathbf{1}^\top A(\theta))$
2. $h(\cdot)$ is weakly differentiable

$$\mathbb{E}\left[\theta_0^\top \widehat{\theta}_\lambda(Y)\right] = -\mathbb{E}\left[\left\langle \frac{h'(Y)}{h(Y)},\ \widehat{\theta}_\lambda(Y)\right\rangle + \sum_i \left(\frac{\partial \widehat{\theta}_{\lambda,i}}{\partial Y_i}(Y)\right)\right]$$

GSURE: unbiased estimator of $\mathbb{E}\left[\left\|\theta_0 - \widehat{\theta}_\lambda\right\|_2^2\right]$

$$\left\|\widehat{\theta}_\lambda\right\|_2^2 + 2\left(\frac{h'(y)}{h(y)}\right)^\top \widehat{\theta}_\lambda + 2\sum_i \left(\frac{\partial \widehat{\theta}_{\lambda,i}}{\partial y_i}(y)\right) + \frac{\mathrm{tr}\ (h''(y))}{h(y)}$$

(Eldar, 2009)

Generalized SURE for continuous exp fam

1. $p_\theta(y) = h(y)\exp(\theta^\top y - \mathbf{1}^\top A(\theta))$
2. $h(\cdot)$ is weakly differentiable

$$\mathbb{E}\left[\theta_0^\top \widehat{\theta}_\lambda(Y)\right] = -\mathbb{E}\left[\left\langle \frac{h'(Y)}{h(Y)},\ \widehat{\theta}_\lambda(Y)\right\rangle + \sum_i \left(\frac{\partial \widehat{\theta}_{\lambda,i}}{\partial Y_i}(Y)\right)\right]$$

GSURE: unbiased estimator of $\mathbb{E}\left[\left\|\theta_0 - \widehat{\theta}_\lambda\right\|_2^2\right]$

$$\left\|\widehat{\theta}_\lambda\right\|_2^2 + 2\left(\frac{h'(y)}{h(y)}\right)^\top \widehat{\theta}_\lambda + 2\sum_i \left(\frac{\partial \widehat{\theta}_{\lambda,i}}{\partial y_i}(y)\right) + \frac{\mathrm{tr}\ (h''(y))}{h(y)}$$

Great, if I want to estimate $\mathbb{E}\left[\left\|\theta_0 - \widehat{\theta}_\lambda\right\|_2^2\right]$

(Eldar, 2009)

Estimating KL

Stein KL Estimator:

$$\widehat{KL}\left(\theta_0 \| \widehat{\theta}_\lambda\right) = \left\langle \widehat{\theta}_\lambda + \frac{h'(y)}{h(y)},\ A'\left(\widehat{\theta}_\lambda\right) \right\rangle + \left\langle A''(\widehat{\theta}_\lambda),\ \frac{\partial \widehat{\theta}_{\lambda,i}}{\partial y_i}(y) \right\rangle - \mathbf{1}^\top A(\widehat{\theta}_\lambda)$$

(Deledalle, 2017)

Estimating KL

Stein KL Estimator:

$$\widehat{KL}\left(\theta_o \| \widehat{\theta}_\lambda\right) = \left\langle \widehat{\theta}_\lambda + \frac{h'(y)}{h(y)},\ A'\left(\widehat{\theta}_\lambda\right) \right\rangle + \left\langle A''(\widehat{\theta}_\lambda),\ \frac{\partial \widehat{\theta}_{\lambda,i}}{\partial y_i}(y) \right\rangle - \mathbf{1}^\top A(\widehat{\theta}_\lambda)$$

with $\mathbb{E}\left[\widehat{KL}\left(\theta_o \parallel \widehat{\theta}_\lambda\right)\right] = KL\left(\theta_o \parallel \widehat{\theta}_\lambda\right) - A(\theta_o)$.

(Deledalle, 2017)

Estimating KL

Stein KL Estimator:

$$\widehat{KL}\left(\theta_o \| \widehat{\theta}_\lambda\right) = \left\langle \widehat{\theta}_\lambda + \frac{h'(y)}{h(y)},\ A'\left(\widehat{\theta}_\lambda\right)\right\rangle + \left\langle A''(\widehat{\theta}_\lambda),\ \frac{\partial \widehat{\theta}_{\lambda,i}}{\partial y_i}(y)\right\rangle - \mathbf{1}^\top A(\widehat{\theta}_\lambda)$$

with $\mathbb{E}\left[\widehat{KL}\left(\theta_o \,\|\, \widehat{\theta}_\lambda\right)\right] = KL\left(\theta_o \,\|\, \widehat{\theta}_\lambda\right) - A(\theta_o)$.

Example: For variance estimation

$$\widehat{KL}\left(\theta_o \,\|\, \widehat{\theta}_\lambda\right) = \frac{1}{4}\left\langle y,\ \widehat{\theta}_\lambda^{-1}\right\rangle + \left\langle \widehat{\theta}_\lambda^{-2},\ \frac{\partial \widehat{\theta}_{\lambda,i}}{\partial y_i}(y)\right\rangle + \frac{1}{2}\mathbf{1}^\top \log(-\widehat{\theta}_\lambda) - \frac{1}{2}$$

(Deledalle, 2017)

Estimating KL

Stein KL Estimator:

$$\widehat{KL}\left(\theta_o \| \widehat{\theta}_\lambda\right) = \left\langle \widehat{\theta}_\lambda + \frac{h'(y)}{h(y)},\ A'\left(\widehat{\theta}_\lambda\right)\right\rangle + \left\langle A''(\widehat{\theta}_\lambda),\ \frac{\partial \widehat{\theta}_{\lambda,i}}{\partial y_i}(y)\right\rangle - \mathbf{1}^\top A(\widehat{\theta}_\lambda)$$

with $\mathbb{E}\left[\widehat{KL}\left(\theta_o \,\|\, \widehat{\theta}_\lambda\right)\right] = KL\left(\theta_o \,\|\, \widehat{\theta}_\lambda\right) - A(\theta_o)$.

Example: For variance estimation

$$\widehat{KL}\left(\theta_o \,\|\, \widehat{\theta}_\lambda\right) = \frac{1}{4}\left\langle y,\ \widehat{\theta}_\lambda^{-1}\right\rangle + \left\langle \widehat{\theta}_\lambda^{-2},\ \frac{\partial \widehat{\theta}_{\lambda,i}}{\partial y_i}(y)\right\rangle + \frac{1}{2}\mathbf{1}^\top \log(-\widehat{\theta}_\lambda) - \frac{1}{2}$$

Solves 1.

(Deledalle, 2017)

The Divergence

Define $\Pi_D = DD^\dagger$, the projection onto $null(D)$.

For TF for Gaussian mean:

$$\widehat{\mathrm{df}}(\widehat{\theta}_\lambda) = \sum_i \frac{\partial \widehat{\theta}_{\lambda,i}}{\partial y_i}(y) = \mathrm{tr}(\Pi_D) = \mathrm{nullity}(D) = \#\ \text{knots} + k + 1$$

(Tibshirani and Taylor, 2012)

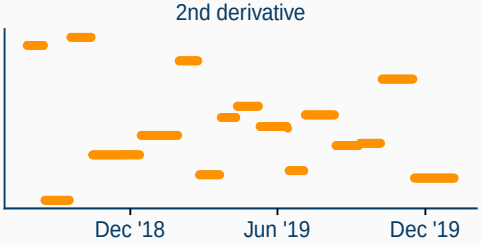
The Divergence

Define $\Pi_D = DD^\dagger$, the projection onto $null(D)$.

For TF for Gaussian mean:

$$\widehat{\mathrm{df}}(\widehat{\theta}_\lambda) = \sum_i \frac{\partial \widehat{\theta}_{\lambda,i}}{\partial y_i}(y) = \mathrm{tr}(\Pi_D) = \mathrm{nullity}(D) = \#\text{ knots} + k + 1$$

Count the pieces $+ k + 1$



(Tibshirani and Taylor, 2012)

Harder case (our result)

For trend filtering with exponential family loss:

$$\frac{\partial \widehat{\theta}_{\lambda,i}}{\partial y_i}(y) = \left(\left(\Pi_D \mathrm{diag}\left(A''(\widehat{\theta}_\lambda) \right) \Pi_D \right)^\dagger \right)_{ii}$$

Harder case (our result)

For trend filtering with exponential family loss:

$$\frac{\partial \widehat{\theta}_{\lambda,i}}{\partial y_i}(y) = \left(\left(\Pi_D \operatorname{diag}\left(A''(\widehat{\theta}_\lambda) \right) \Pi_D \right)^\dagger \right)_{ii}$$

- Solves 2.
- Variance estimation: $A''(\theta) = \frac{1}{2\theta^2}$
- Final form:

$$\widehat{KL}\left(\theta_0 \,\|\, \widehat{\theta}_\lambda \right) = -\frac{1}{2} + \sum_i \frac{y_i}{4\widehat{\theta}_{\lambda,i}} + \frac{2\left(\left(\Pi_D \operatorname{diag}\left(\widehat{\theta}_\lambda^{-2} \right) \Pi_D \right)^\dagger \right)_{ii}}{\widehat{\theta}_{\lambda,i}^2} + \frac{\log(-\widehat{\theta}_{\lambda,i})}{2}$$

Benefits

+ Measures the curvature correctly

+ No sample splitting, recomputing

+ Fairly interpretable

+ Estimating the risk we control theoretically

– Π_D not nearly as clean

Theory

Convergence result

1. n is large enough
2. λ_n is large enough to control the empirical process
3. θ_o is k-times differentiable, and $\mathrm{TV}(\theta_o^{(k)}) < C_n$
4. Observations on a d-dimensional regular grid
5. Ignore log factors which are myriad and ugly

Theorem:

$$\frac{1}{n}\mathrm{KL}\left(\theta_o \,\|\widehat{\theta}_{\lambda_n}\right) = \begin{cases} O_p\left(\left(\frac{1}{n}\right)^{\frac{k+1}{d}}\right) & d \geq 2k+2 \\ O_p\left(\left(\frac{1}{n}\right)^{\frac{2k+2}{2k+2+d}}\right) & d < 2k+2 \end{cases}$$

Notes on our theorem

$$\frac{1}{n}\mathrm{KL}\left(\theta_{\mathrm{o}} \,\|\widehat{\theta}_{\lambda_n}\right) = \begin{cases} O_p\left(\left(\frac{1}{n}\right)^{\frac{k+1}{d}}\right) & d \geq 2k+2 \\ O_p\left(\left(\frac{1}{n}\right)^{\frac{2k+2}{2k+2+d}}\right) & d < 2k+2 \end{cases}$$

- − Our log factors are worse than for (sub)-Gaussian case
- − Our log factors are worse than some tailored proofs elsewhere
- \+ Ignoring log factors, this is minimax optimal

see also Sadhanala et al. (2017)

Sketch of proof

- Can use properties of exponential families to get “Basic inequality”

$$KL\left(\theta_0 \,\|\, \widehat{\theta}\right) \leq (Y - A'(\theta_0))^\top(\theta_0 - \widehat{\theta}) + \lambda \|D\theta_0\| - \lambda \left\|D\widehat{\theta}\right\|$$

- First term is empirical process, second term controlled by λ
- $Y - A'(\theta_0)$ is mean zero, sub-exponential
- Play some games

Sketch of proof

- Can use properties of exponential families to get “Basic inequality”

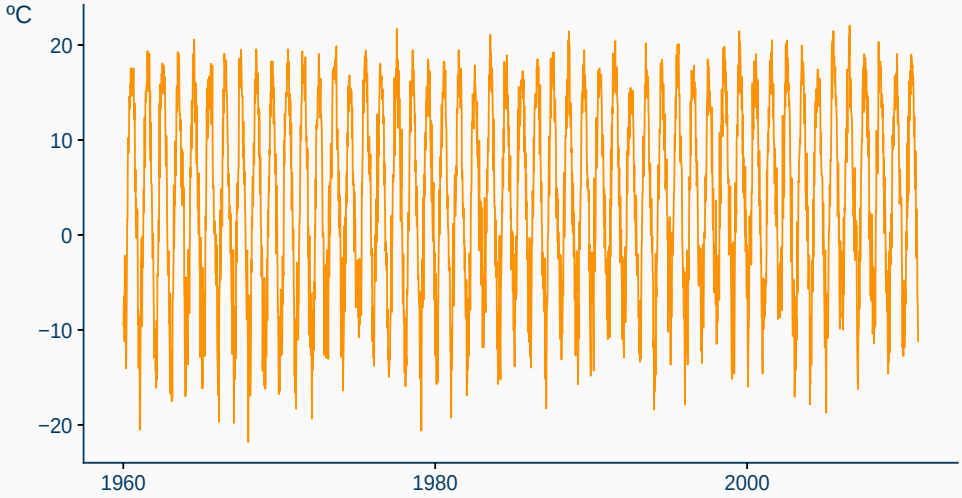
$$KL\left(\theta_0 \,\|\, \widehat{\theta}\right) \leq (Y - A'(\theta_0))^\top(\theta_0 - \widehat{\theta}) + \lambda \|D\theta_0\| - \lambda \left\|D\widehat{\theta}\right\|$$

- First term is empirical process, second term controlled by λ
- $Y - A'(\theta_0)$ is mean zero, sub-exponential
- Play some games

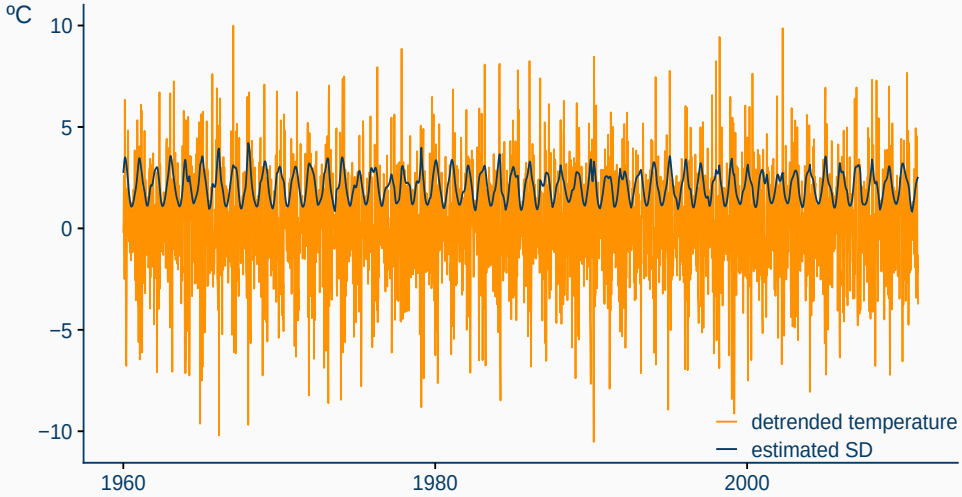
. . . going on 15 pages of LaTeX. . .

Empirical results

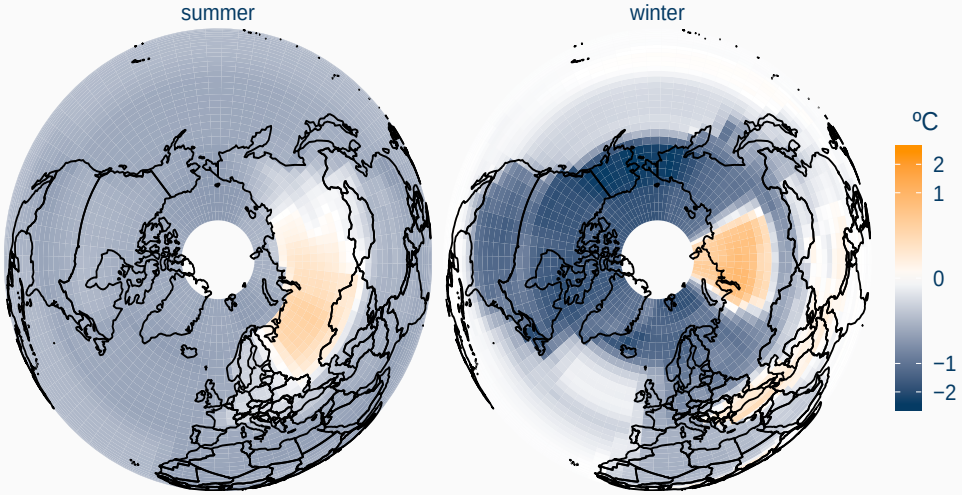
Toronto temperature



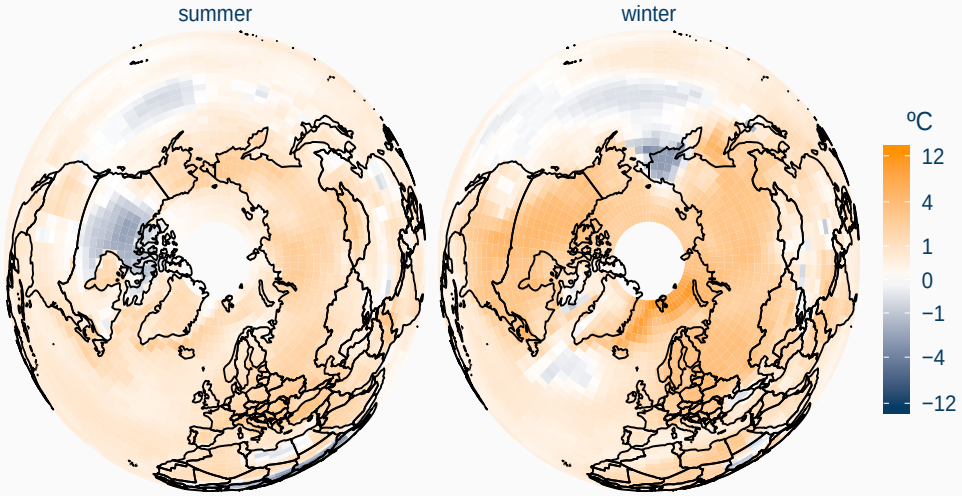
Toronto temperature



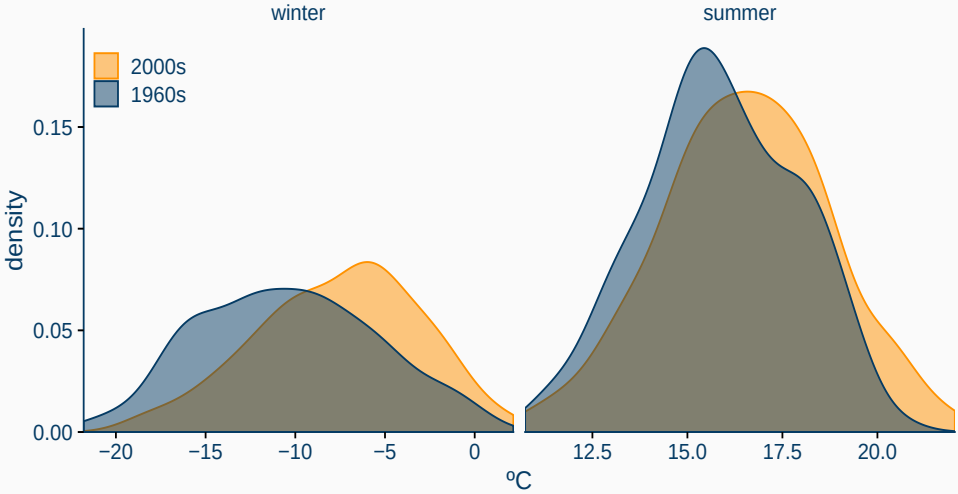
Change in estimated SD (1960s vs 2000s)



Change in mean temperature (1960s vs 2000s)



Observed temperatures in Toronto (1960s vs 2000s)



Conclusion

Wrapping up

We generalized TF to exponential families

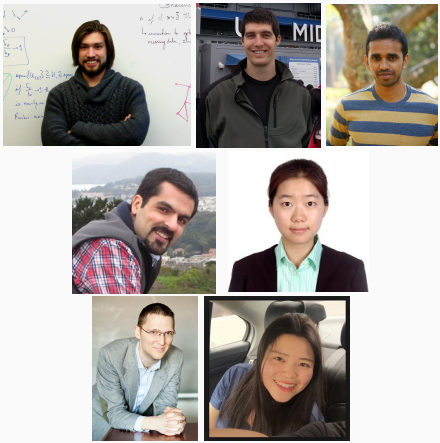
- Developed tailored algorithms for some big data
- Derived risk estimator to select λ w/o excess computation
- Proved theory for nonparametric function estimation

Future work

- Do we care about θ? $A'(\theta)$?
- Multiparameter exponential families?
- Model selection in discrete case?
- TF shrinks the estimate. Maybe reestimate using learned knots?
- Model misspecification relative to the actual data

Real MODIS track

Collaborators and funding



Institute for
New Economic Thinking

Appendix

Detailed PDIP

Primal	Dual
$\min_{\theta} \quad f(\theta) + \lambda \left\| D\theta \right\|_1$	$\min_{v} \quad f^*(-D^\top v)$
	s.t. $\quad \left\| v \right\|_\infty \leq \lambda$

- $f(\theta) := \sum \theta_i + y_i e^{-\theta_i}$
- $f^*(u) := \sum (u_i - 1) \log \frac{y_i}{1-u_i} + u_i - 1$

Perturbed KKT conditions ($w > 0$) $\Longrightarrow$

$$r_w(v, \mu_1, \mu_2) := \begin{bmatrix} \nabla f^*(-D^\top v) + D(v - \lambda \mathbf{1})^\top \mu_1 - D(v + \lambda \mathbf{1})^\top \mu_2 \\ -\mu_1 (v - \lambda \mathbf{1}) + \mu_2 (v + \lambda \mathbf{1}) - w^{-1} \mathbf{1} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

- As $w \longrightarrow \infty$, this converges to the optimum.
- But this is a nonlinear system, can't solve.
- Use Newton steps, which give the $[Ms = v]$ thing
- M is the Jacobian of r_w.

Locally adaptive regression splines

$$\min_{f \in \mathcal{F}_k} \frac{1}{2n} \|y - f\|_2^2 + \lambda \mathrm{TV}(f^{(k)})$$

- $\mathcal{F}_k = \left\{ f : [0,1] \to \mathbb{R},\ f^{(k)} \text{ exists a.e.},\ TV\left(f^{(k)}\right) < \infty \right\}$
- Solution is a k^{th}-degree spline (Mammen and van de Geer, 1997)
- $k \geq 2$ knots are not generally at the input points
- Not generically computable, but a close relative is (whose knots are at the inputs)
- Solve

$$\min_{\theta} \frac{1}{2n} \|y - G\theta\|_2^2 + \lambda \|C\theta\|_1$$

- Either G or C dense, $(n \times n)$.

Smoothing splines

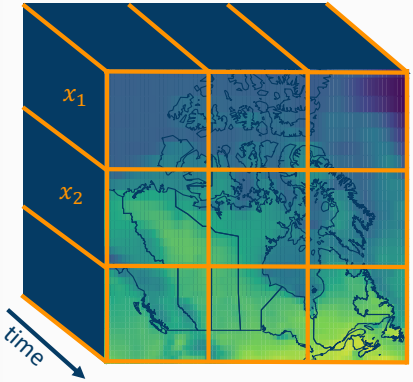
$$\min_{f \in \mathcal{W}_{(k+1)/2}} \frac{1}{2n} \|y - f\|_2^2 + \lambda \int_{\mathcal{X}} \left(f^{\left(\frac{k+1}{2}\right)}(t) \right)^2 dt$$

- $\mathcal{W}_{(k+1)/2)} = \left\{ f : [0,1] \longrightarrow \mathbb{R},\ f^{(k)} \text{ exists},\ \int_{\mathcal{X}} \left(f^{\left(\frac{k+1}{2}\right)}(t) \right)^2 dt < \infty \right\}$
- Solution is a k^{th}-degree spline (Wahba, 1990)
- k needs to be odd
- One way to solve:

$$\min_{\theta} \frac{1}{2n} \|y - \theta\|_2^2 + \lambda \|K\theta\|_1$$

- K is banded, so solution requires $O(n)$ computations.

Consensus version



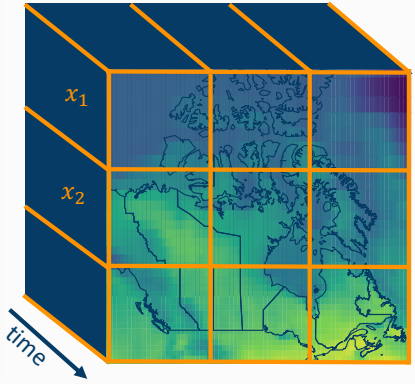
$$\min_{x_g=\theta\ \forall g} \sum_{g\in G} -\ell(x_g) + \lambda \left\| D_g . x_g \right\|_1$$

$$x_g \leftarrow \underset{x_g}{\operatorname{argmin}} -\ell(x_g) + \lambda \left\| D_g . x_g \right\|_1 + u^\top (x_g - \theta) + \frac{\rho}{2} \left\| x_g - \theta \right\|_2^2$$

$$\theta \leftarrow \operatorname{avg}(x_g + u_g/\rho)$$

$$u_g \leftarrow u_g + \rho(x_g - \theta)$$

Consensus version



$$\min_{x_g=\theta\ \forall g} \sum_{g\in G} -\ell(x_g) + \lambda \left\| D_g . x_g \right\|_1$$

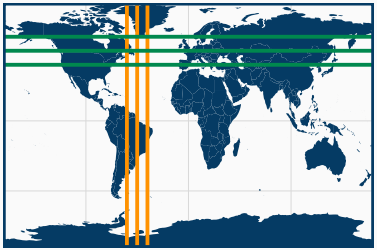
$$x_g \leftarrow \underset{x_g}{\operatorname{argmin}} -\ell(x_g) + \lambda \left\| D_g . x_g \right\|_1 + u^\top (x_g - \theta) + \frac{\rho}{2} \left\| x_g - \theta \right\|_2^2$$

$$\theta \leftarrow \operatorname{avg}(x_g + u_g/\rho)$$

$$u_g \leftarrow u_g + \rho(x_g - \theta)$$

Requires very few iterations, but iterations are $O\left(|\text{block}|^3\right)$. Can parallelize over blocks.

Grid world



$$\min_{\theta=a=b=c} \sum_{ijt} -\ell(\theta_{ijt}) + \lambda \sum_{it} \|Da_{i\cdot t}\|_1 + \lambda \sum_{jt} \|Db_{\cdot jt}\|_1 + \lambda \sum_{ij} \|Dc_{ij\cdot}\|_1$$

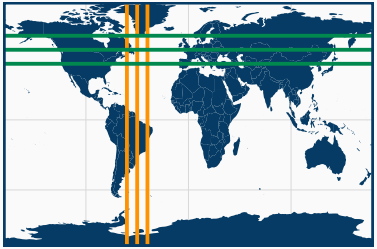
$$\theta_{ijt} \longleftarrow \text{solution of } A'(\theta_{ijt}) = k^{(1)}_{ijt}\theta_{ijt} + k^{(2)}_{ijt}$$

$$[a, b, c] \longleftarrow \mathrm{TF}_{1d}([a, b, c] + [u, v, w])$$

$$[u, v, w] \longleftarrow [u, v, w] + \theta - [a, b, c]$$

$$k^{(1)}, k^{(2)} \longleftarrow \text{simple linear functions of } a, b, c, u, v, w$$

Grid world



$$\min_{\theta=a=b=c} \sum_{ijt} -\ell(\theta_{ijt}) + \lambda \sum_{it} \|Da_{i\cdot t}\|_1 + \lambda \sum_{jt} \|Db_{\cdot jt}\|_1 + \lambda \sum_{ij} \|Dc_{ij\cdot}\|_1$$

$$\theta_{ijt} \leftarrow \text{solution of } A'(\theta_{ijt}) = k^{(1)}_{ijt}\theta_{ijt} + k^{(2)}_{ijt}$$

$$[a, b, c] \leftarrow \mathrm{TF}_{1d}([a, b, c] + [u, v, w])$$

$$[u, v, w] \leftarrow [u, v, w] + \theta - [a, b, c]$$

$$k^{(1)}, k^{(2)} \leftarrow \text{simple linear functions of } a, b, c, u, v, w$$

Requires many iterations, but iterations are $O(|\text{line}|)$. Can parallelize over lines.

Which classes and cannonical scaling

- D is such that it smooths over axis parallel lines in the grid
- Define $\mathcal{K}_d^k(C_n) = \{\theta : \|D\theta\|_1 < C_n\}$
- Define $\mathcal{H}_d^{k+1}(L)$ to be the Hölder class containing discretized Hölder smooth-functions with k derivatives
- Can show that $\mathcal{H}_d^{k+1}(L) \subset \mathcal{K}_d^k(cLn^{1-(k+1)/d})$
- This gives the lower bound.
- Linear smoothers can't achieve this rate (Donoho and Johnstone, 1998)

Confidence intervals

Like LASSO other ℓ_1-regularized methods, this is biased

Full Hessian at the solution would be insane

Marginal coverage could be done numerically (but the bias)

One approach would be “relaxed” TF

(Very) recent work uses this for LASSO CIs

Ongoing work with Max Ferrell at Chicago Booth

Also, how does the (known) bias compare to the (unknown) misspecification

Sources of misspecification

Real satellite track

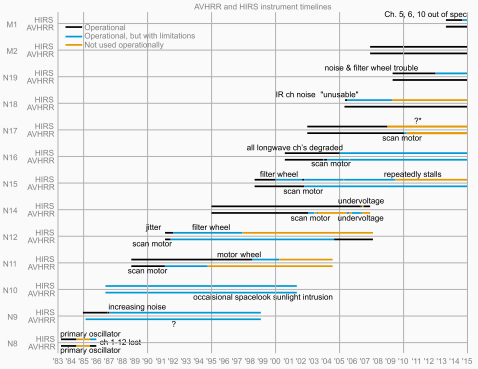
Track overlap

Angular distortion of instruments

Degradation of instrument quality (theoretically, more in mean than variance)

Intersatellite calibration

Data interpolation from AVHRR and HIRS



Source: (Staten et al., 2016)

Selected references

Deledalle, C.-A. (2017), "Estimation of Kullback-Leibler losses for noisy recovery problems within the exponential family," *Electronic Journal of Statistics*, **11**, 3141–3164.

Donoho, D. L., and Johnstone, I. M. (1998), "Minimax estimation via wavelet shrinkage," *The Annals of Statistics*, **26**(3), 879–921.

Efron, B. (1986), "How biased is the apparent error rate of a prediction rule?" *Journal of the American Statistical Association*, **81**(394), 461–470.

Eldar, Y. C. (2009), "Generalized SURE for exponential families: Applications to regularization," *IEEE Transactions on Signal Processing*, **57**, 471–481.

Green, P. J., and Silverman, B. W. (1994), *Nonparametric regression and generalized linear models: a roughness penalty approach*, Chapman and Hall/CRC, Boca Raton, FL.

Hütter, J.-C., and Rigollet, P. (2016), "Optimal rates for total variation denoising," in *29th Annual Conference on Learning Theory*, eds. V. Feldman, A. Rakhlin, and O. Shamir, vol. 49 of *Proceedings of Machine Learning Research*, pp. 1115–1146, Columbia University, New York, New York, USA, PMLR.

Khodadadi, A., and McDonald, D. J. (2019), "Algorithms for estimating trends in global temperature volatility," in *Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI-19)*, eds. P. V. Hentenryck and Z.-H. Zhou, Association for the Advancement of Artificial Intelligence.

Kim, S.-J., Koh, K., Boyd, S., and Gorinevsky, D. (2009), "ℓ_1 trend filtering," *SIAM Review*, **51**(2), 339–360.

Mammen, E., and van de Geer, S. (1997), "Locally adaptive regression splines," *The Annals of Statistics*, **25**(1), 387–413.

Sadhanala, V. (2019), "Nonparametric methods with total variation type regularization," Ph.D. thesis, Carnegie Mellon University.

Sadhanala, V., Wang, Y.-X., Sharpnack, J. L., and Tibshirani, R. J. (2017), "Higher-order total variation classes on grids: Minimax theory and trend filtering methods," in *Advances in Neural Information Processing Systems 30*, eds. I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, pp. 5800–5810, Curran Associates, Inc.

Selected references

Staten, P. W., Kahn, B. H., Schreier, M. M., and Heidinger, A. K. (2016), "Subpixel characterization of HIRS spectral radiances using cloud properties from AVHRR," *Journal of Atmospheric and Oceanic Technology*, **33**(7), 1519–1538.

Stein, C. M. (1981), "Estimation of the mean of a multivariate normal distribution," *The Annals of Statistics*, **9**(6), 1135–1151.

Tibshirani, R. J. (2014), "Adaptive piecewise polynomial estimation via trend filtering," *Annals of Statistics*, **42**, 285–323.

Tibshirani, R. J., and Taylor, J. (2012), "Degrees of freedom in lasso problems," *Annals of Statistics*, **40**, 1198–1232.

Vaiter, S., Deledalle, C., Fadili, J., Peyré, G., and Dossal, C. (2017), "The degrees of freedom of partly smooth regularizers," *Annals of the Institute of Statistical Mathematics*, **69**, 791–832.

Wahba, G. (1990), *Spline models for observational data*, vol. 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA.

Wang, Y.-X., Sharpnack, J., Smola, A. J., and Tibshirani, R. J. (2016), "Trend filtering on graphs," *Journal of Machine Learning Research*, **17**(105), 1–41.