Trend filtering in exponential families

Daniel J. McDonald

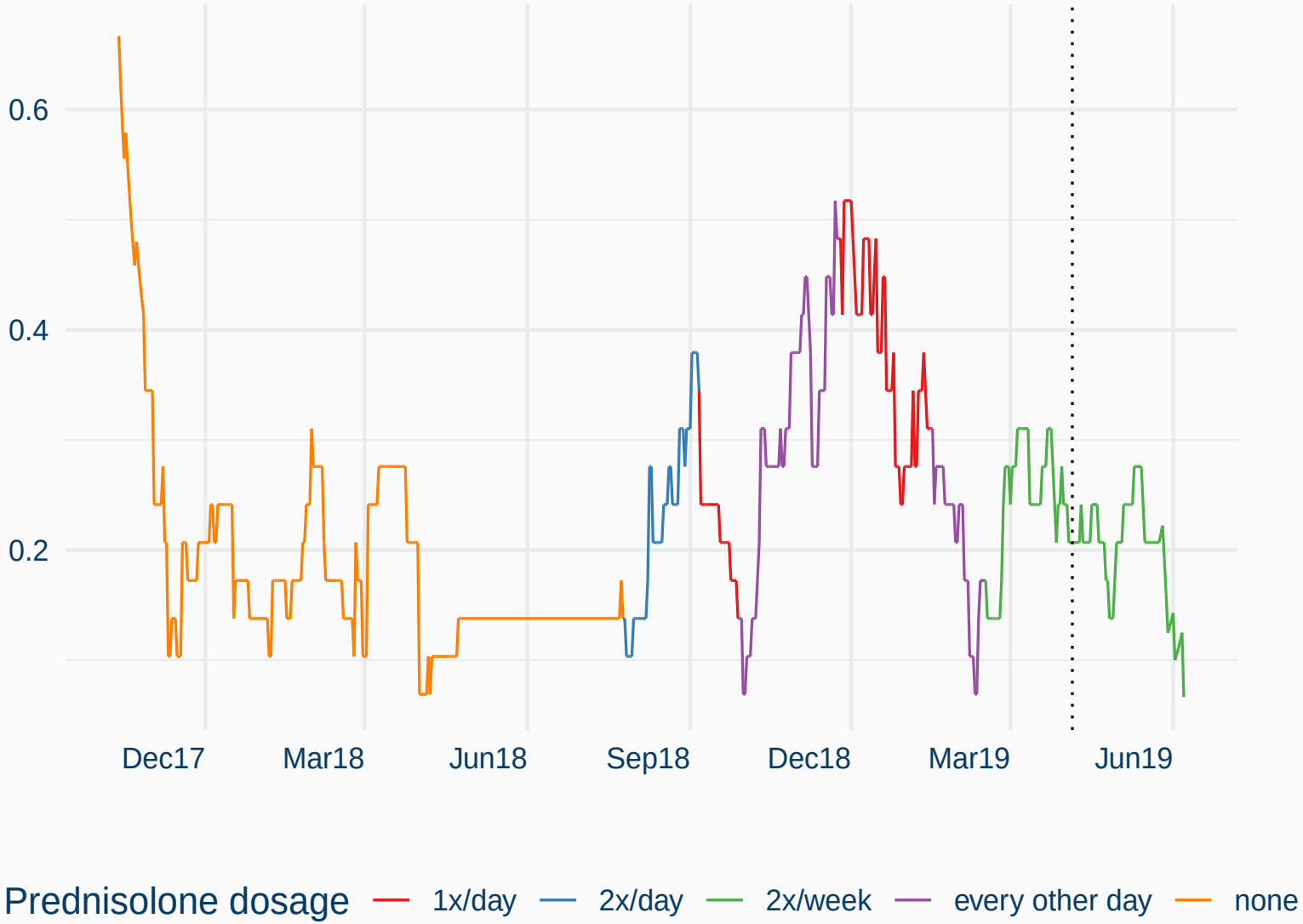
Indiana University, Bloomington

pages.iu.edu/~dajmcdon

29 May 2019

These are my cats

Number of vomits/day (30 day rolling average)



Poisson model

y_i is the number of vomits on day i

Poisson distributed with time-varying parameter ϕ_i

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

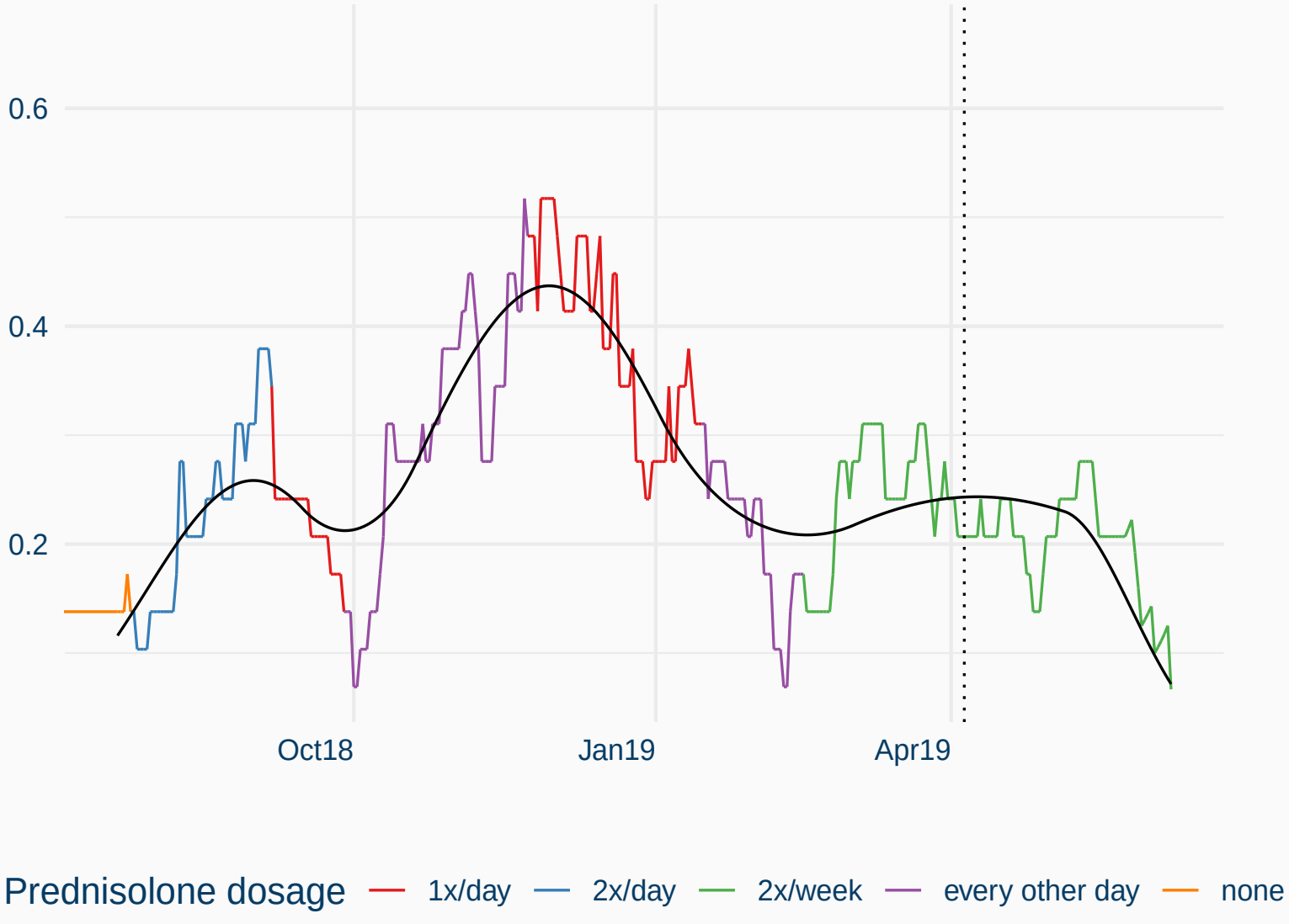
Goal: estimate ϕ from data, ϕ should be "smooth".

Set $\theta_i = \log \phi_i$

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

D matrix enecodes smoothness

Trend filtering



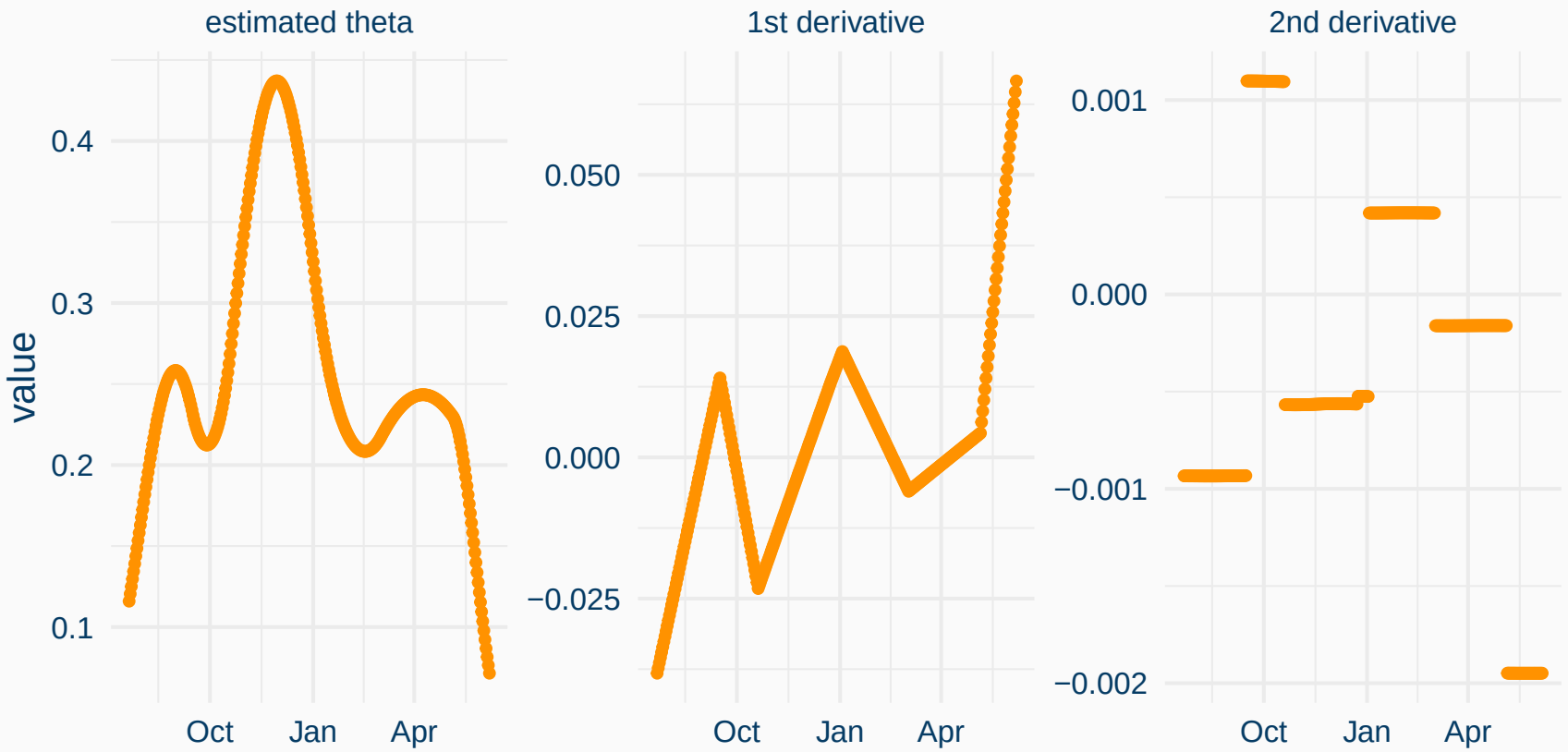
Second-order Poisson trend filtering

$$D = \begin{bmatrix} 1 & -2 & 1 & & 0 \\ & & \ddots & & \\ 0 & & 1 & -2 & 1 \end{bmatrix} \in \mathbb{R}^{(n-2)\times n}$$

Looks visually like a smoothing spline, but more locally adaptive

Works well on functions of "bounded variation": $\int_a^b |\theta''(x)|dx < \infty$

Derivative properties



What's this talk about

Trend filtering is not new.

But, aside from small specializations, the theory/methods are for additive (sub)-Gaussian noise only.

1. Motivated by spatio-temporal variance estimation from weather satellites.
2. We generalize to exponential families.
3. Provide some algorithms that work on big data.
4. Provide justifiable way of selecting the tuning parameter.

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

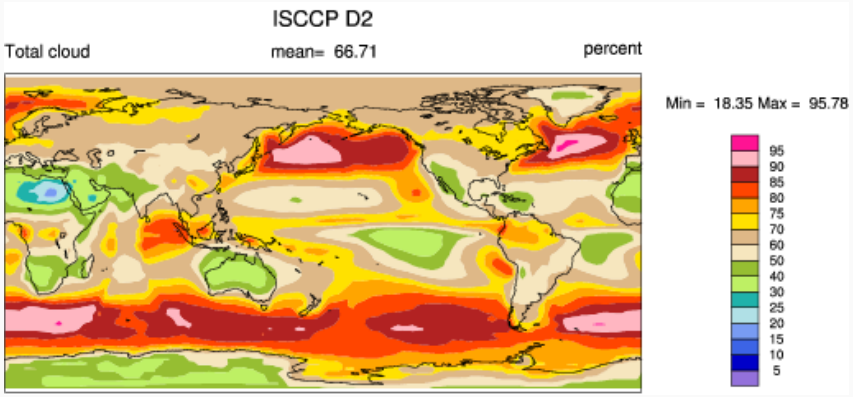
Increasing mean temperature understates the costs:

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

Using weather satellites

Drivers of climate variation:

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

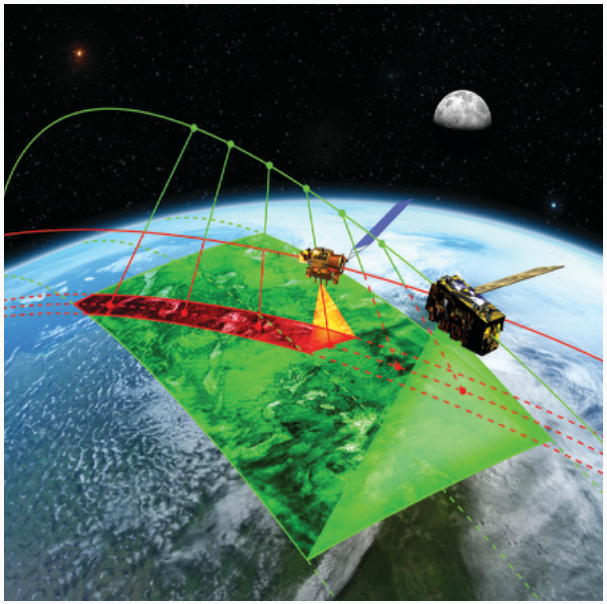


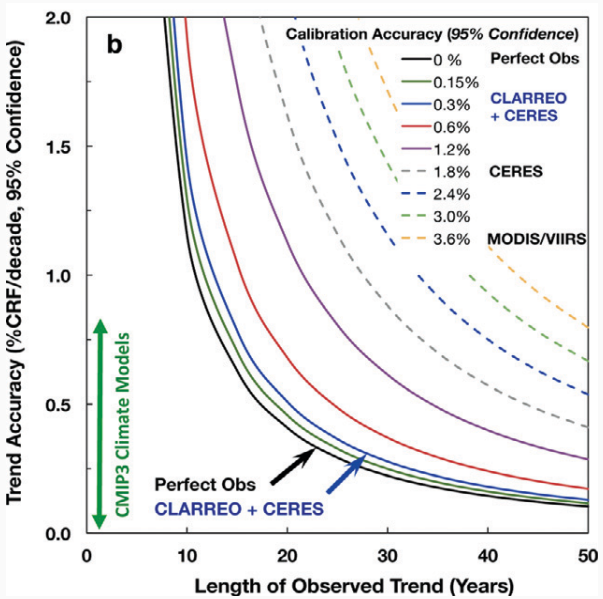
CLARREO satellite: monitor cloud top temperature as it relates to climate.

- Has yet to launch, no sooner than 2022
- Defunded in most recent federal budget

Source: NCAR CCSM3 Diagnostic Plots.

CLARREO vs MetOp/Modis





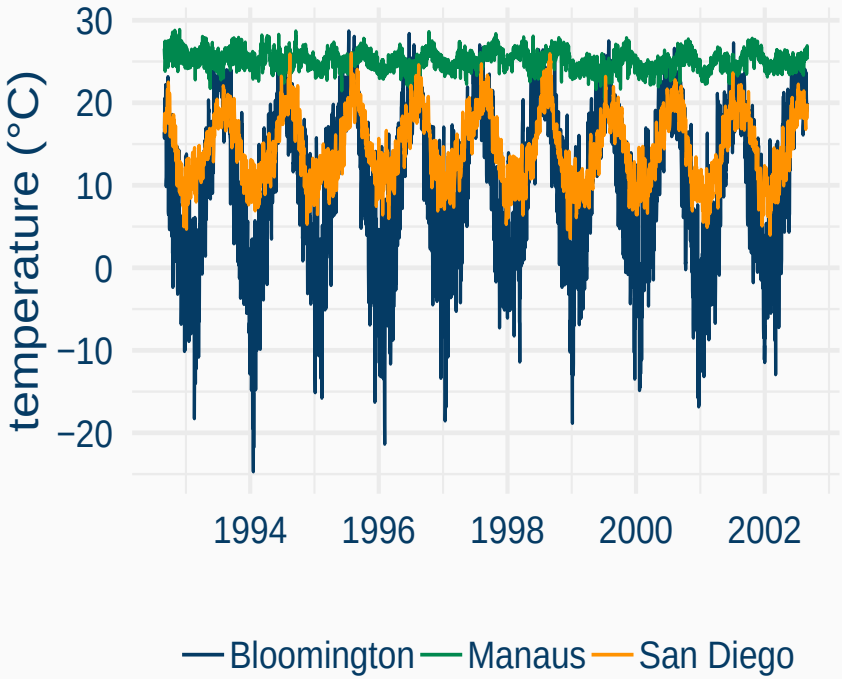
- Weather satellites aren't made for this.
- More information in higher moments than in average?

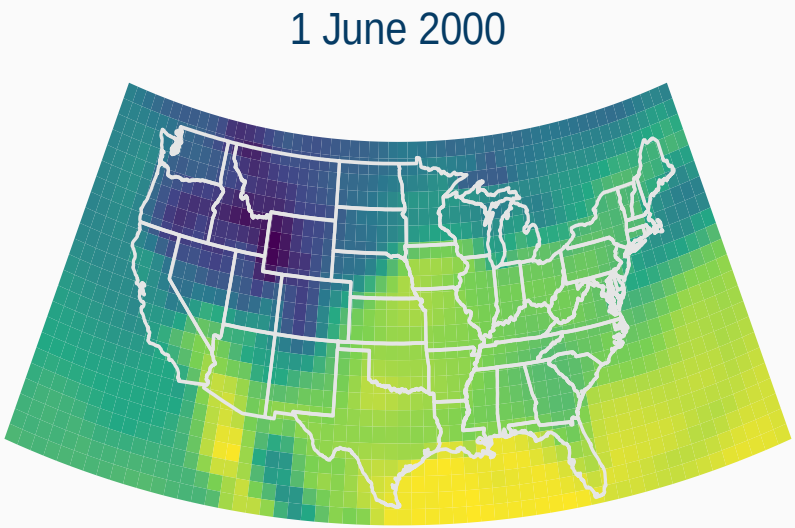
Source: Wielicki, et al. (2013).

Satellite data

Once collaborators do lots of processing…

- 52,000 time series
- daily records over ~ 40 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 $x_{ijt} \sim \text{Normal}\left(0, \sigma^2_{ijt}\right)$
- Estimate σ^2, but it should be "smooth" relative to space and time.
- Use a matrix D + penalty to encode this smoothness.

Exponential families

Natural exponential family

Fix a (σ-finite) measure μ on the Borel subsets of $\mathbb{R}$

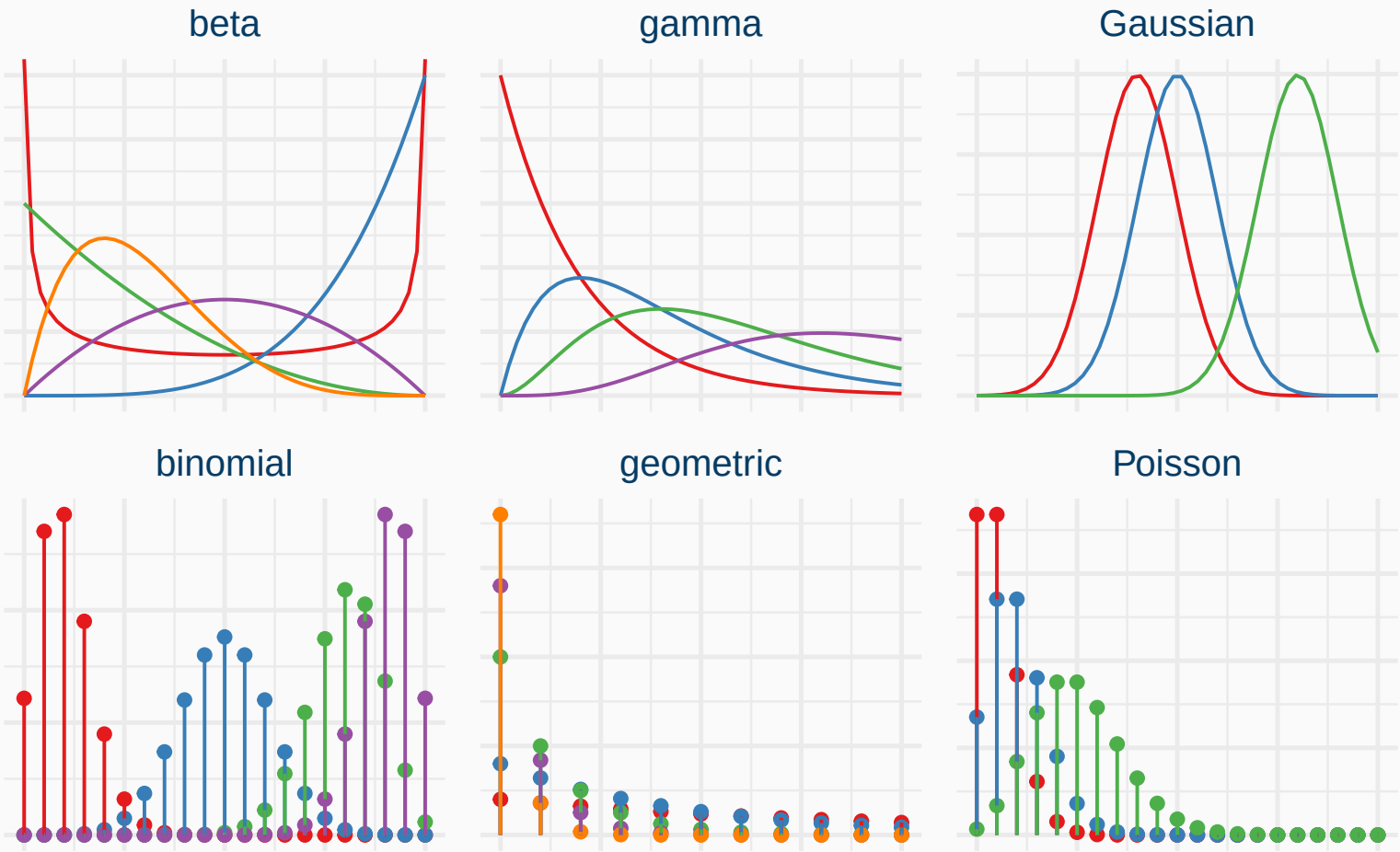
Let $\Theta = \left\{\theta : \int d\mu(y) exp(y\theta) < \infty\right\}$.

Let $A(\theta) = \log \int d\mu(y) exp(y\theta)$.

Then $p_\theta(y) = \exp(y\theta - A(\theta))$ is an exponential family w.r.t. μ.

When convenient, write $p_\theta(y) = h(y) \exp(y\theta - A(\theta))$, for base measure h

Standard examples



Some important properties we will use

Let $\theta_0 \in \Theta$ be in the interior.

1. Θ is convex, A is strictly convex on Θ
2. All derivatives of A exist at θ_0.
3. $\mathbb{E}[y] = A'(\theta_0)$, $\mathbb{V}[y] = A''(\theta_0)$.
4. The cumulant generating function is $A(\theta_0)$.
5. $KL(\theta_1 \,||\, \theta_2) = A(\theta_2) - A(\theta_1) + (\theta_1 - \theta_2)A'(\theta_1)$
6. y is sub-exponential: $\mathbb{P}(y > \epsilon) < c\exp(-\alpha\epsilon)$, some α, c.

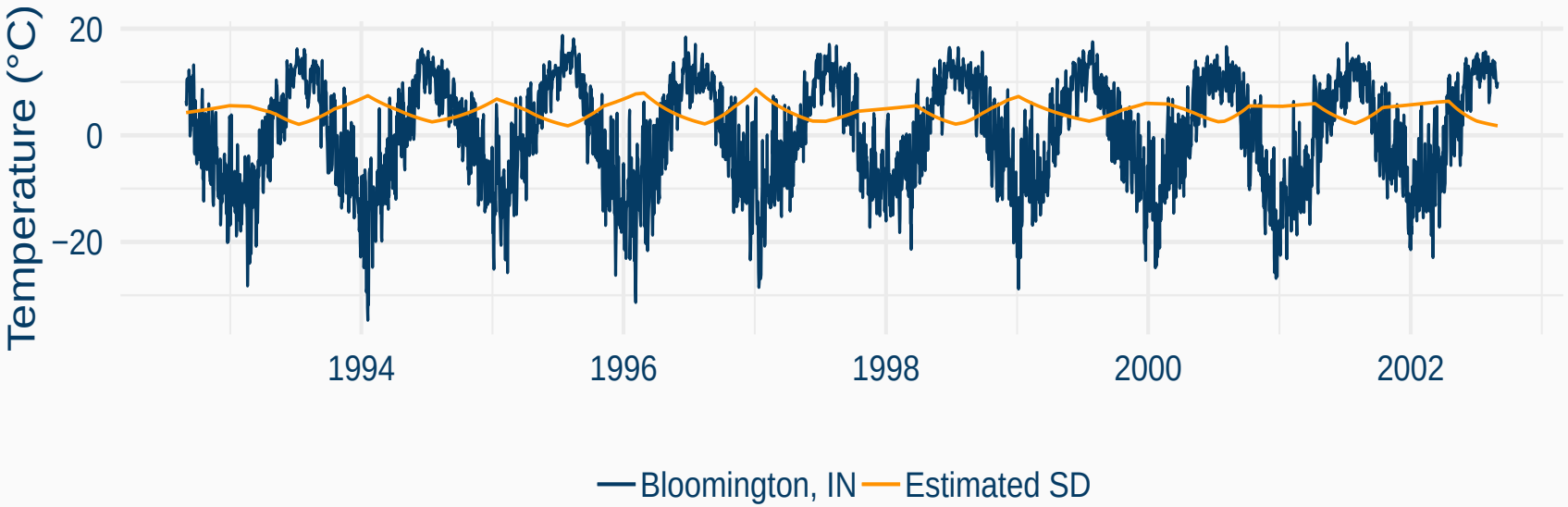
Algorithms

Optimization problem

$$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2}\right) \Longrightarrow \frac{1}{2\sqrt{\pi y}} \exp\left(y\theta - A(\theta)\right) \quad A(\theta) := -\frac{1}{2}\log(-\theta)$$

$$\min_{\theta} \sum_{ijt} A(\theta_{ijt}) - y_{ijt}\theta_{ijt} + \lambda \left\|D\theta\right\|_1$$

Standard optimizer: Primal Dual Interior Point method.



Generic PDIP

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

The matrix $M = M\left(\theta^{(k)}\right)$, dense, and roughly $10^9 \times 10^9$.

This isn't going to work.

Alternating direction method of multipliers

One way to solve optimization problems like this is to 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 with $\rho > 0$

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

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

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

Why would you do this?

- It decouples f and g: this can be easier
- If f and g are nice, the updates can be parallelized
- The algorithm converges under very general conditions
- There are often many ways to decouple a problem

$$\min_{\theta} -\ell(\theta) + \lambda \|D\theta\|_1$$

- The individual minimizations don't have to be solved in closed form

Example:

$$\theta \leftarrow \operatorname*{argmin}_{\theta} -\ell(\theta) + \frac{\rho}{2} \|D\theta - \alpha + u\|_2^2$$

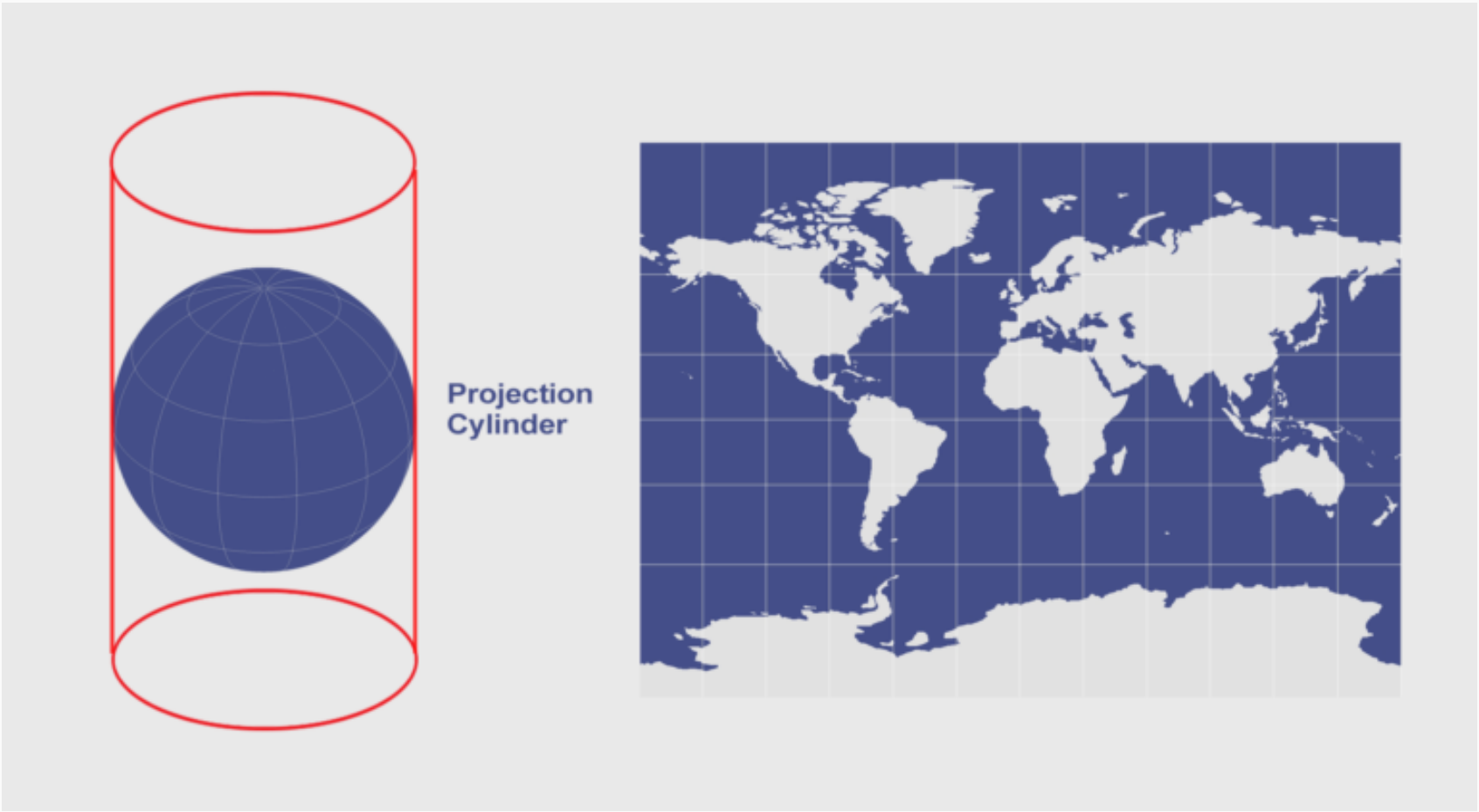
$$\alpha \leftarrow \mathcal{S}_{\lambda/\rho}(D\theta + u)$$

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

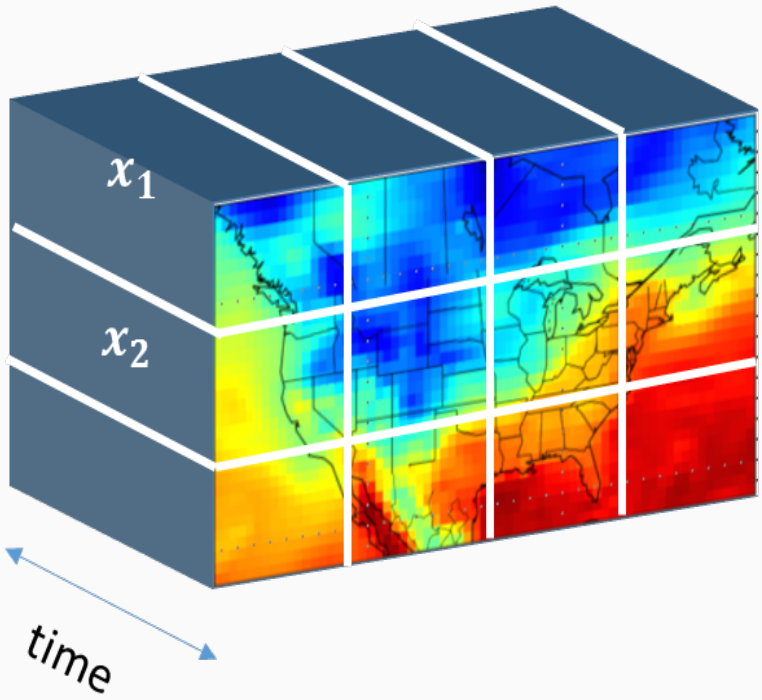
$$[\mathcal{S}_a(b)]_k = \operatorname{sgn}(b_k)(|b_k| - a)_+$$

Real MODIS track

What our data look like



Concensus version



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

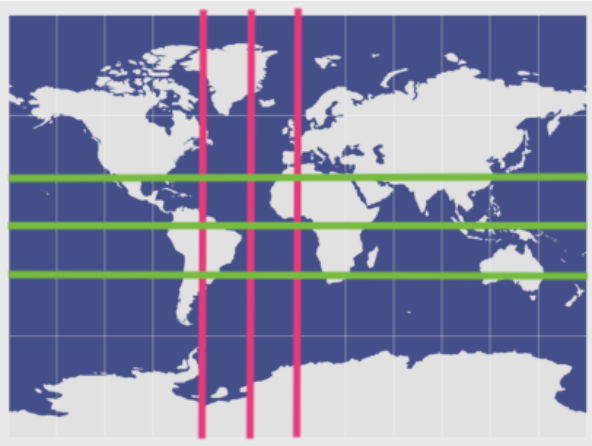
$$x_g \leftarrow \operatorname*{argmin}_{x_g} -\ell(x_g) + \lambda \left\|D_g.x_g\right\|_1 + u^\top(x_g-\theta) + \frac{\rho}{2}\left\|x_g-\theta\right\|_2^2$$

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

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

Requires very few iterations, but iterations are $O(n^3)$. Can parallelize over blocks.

Grid world



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

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

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

$$(u, v, w) \leftarrow (u, v, w) + \theta - (a, b, c)$$

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

Requires many iterations, but iterations are $O(n)$. Can parallelize over lines.

$$\mathrm{TF}_{1d}(z) := \mathrm{argmin}_x \|x - z\|_2^2 + \lambda/\rho \|Dx\|_1$$

Hints and caveats

- Linearized ADMM if large memory computer
- Can come up with intermediate options
- Off-the-shelf stuff doesn’t work
- Smaller problems don’t need these details
- Must repeat for many tuning parameters

Tuning parameter selection

Risk

- Suppose you have observed data Y, a predictor $f_\lambda(Y)$
- You want to know $MSE(\lambda) = \mathbb{E}\left[\|Y - f_\lambda(Y)\|_2^2\right]$
- Examining $Error(\lambda) = \|Y - f_\lambda(Y)\|_2^2$ is biased if you used Y to get f
- AIC, BIC, GCV compensate with $Error(\lambda) + pen(\lambda)$
- Cross Validation uses held-out sets

Unbiased estimation

- If data are i.i.d., noise is additive, mean zero

$$MSE(\lambda) = \mathbb{E}\left[\|Y - f_\lambda(Y)\|_2^2\right] - n\sigma^2 + 2\text{tr Cov}(Y, f_\lambda(Y))$$

- If $f_\lambda = WY$, then $\text{tr Cov}(Y, f_\lambda(Y)) = \text{tr } W =: \sigma^2 df(\lambda)$
- If $Y \sim \text{Normal}(\mu, \sigma^2 I_n)$ and f weakly differentiable with ess. bounded partials, then

$$\text{tr Cov}(Y, f_\lambda(Y)) = \sigma^2 \mathbb{E}\left[\text{tr } \frac{\partial f_\lambda(y)}{\partial y_i}\bigg|_Y\right]$$

- Ingredients for SURE:
  1. Expression for risk I want, w/o dependence on parameters
  2. Expression for $\text{tr } \frac{\partial f_\lambda(y)}{\partial y_i}$

SURE for continuous exp fam

1. $p_\theta(y) = h(y)\exp(\theta^\top y - A(\theta))$
2. f weakly differentiable with ess. bounded partials
3. h is weakly differentiable

$$\mathbb{E}\left[\theta^\top f(Y)\right] = -\mathbb{E}\left[\left(\frac{\nabla h(Y)}{h(Y)}\right)^\top f(Y) + \operatorname{tr}\frac{\partial f_\lambda(y)}{\partial y_i}\Big|_Y\right]$$

- Conincides with previous for Gaussian case
- Can be used to get unbiased estimator of $\mathbb{E}\left[\left\|\theta - \widehat{\theta}\right\|_2^2\right]$

Estimating KL

Result (Deledalle '17)

$$\widehat{KL}(\widehat{\theta} \,\|\, \theta_o) = \left\langle \widehat{\theta} + \frac{\nabla h(Y)}{h(Y)},\, A'(\widehat{\theta}) \right\rangle + (A''(\widehat{\theta}))^\top \left(\frac{\partial \widehat{\theta}_i}{\partial y_i}\Big|_Y \right) - 1^\top A(\widehat{\theta}),$$

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

Here:

$$\widehat{KL}(\widehat{\theta} \,\|\, \theta_o) = \sum_{ijt} \left(\frac{Y_{ijt}}{4\widehat{\theta}_{ijt}} + \frac{1}{2\widehat{\theta}_{ijt}^2} \frac{\partial \widehat{\theta}_{ijt}}{\partial y_{ijt}}\Big|_Y + \frac{\log(-\widehat{\theta}_{ijt})}{2} \right) - \frac{n}{2}$$

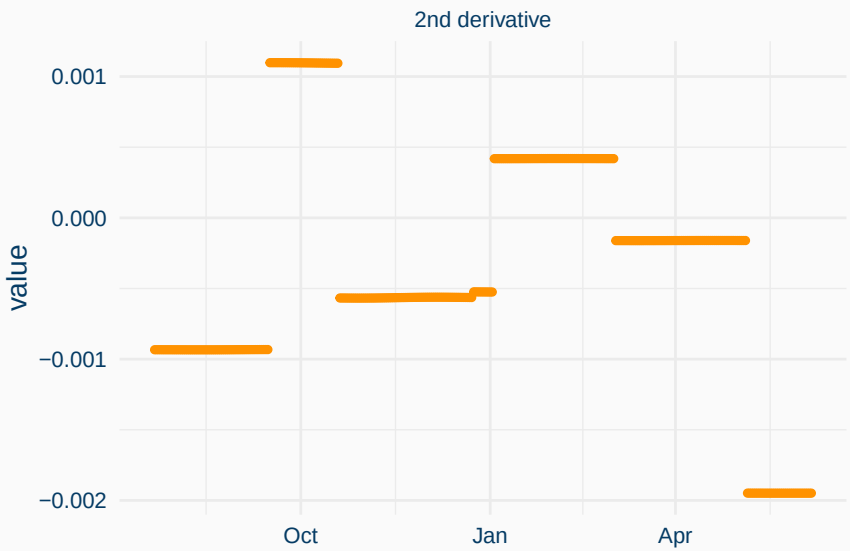
The Divergence

Let G be the rows of D with $D\widehat{\theta} = 0$.

Define $\Pi_G = GG^{\dagger}$, the projection onto $null(G)$.

For Trend-Filtering with Gaussian loss (Tibshirani and Taylor '14):

$$\mathrm{df}(\widehat{\theta}) = \sum_i \frac{\partial \widehat{\theta}_i}{\partial y_i} = \mathrm{tr}(\Pi_G) = \mathrm{nullity}(G).$$



- Return of the cats
- All you have to do is count the number of pieces.

Harder case

- For trend filtering with exponential family loss (Jacobian):

$$\frac{\partial \widehat{\theta}_i}{\partial y_i} = \operatorname{diag}\left(\Pi_G \left(\Pi_G \operatorname{diag}\left(A''(\widehat{\theta})\right) \Pi_G\right)^{\dagger} \Pi_G\right)$$

- Our case: $A''(\theta) = \dfrac{1}{2\theta^2}$
- Final form: $\widehat{KL}(\widehat{\theta} \,\|\, \theta_0) = \displaystyle\sum_{ijt} \left(\frac{Y_{ijt}}{4\widehat{\theta}_{ijt}} + \frac{1}{4\widehat{\theta}_{ijt}^4} + \frac{\log(-\widehat{\theta}_{ijt})}{2} \right) - \frac{n}{2}$

Theory

Briefly

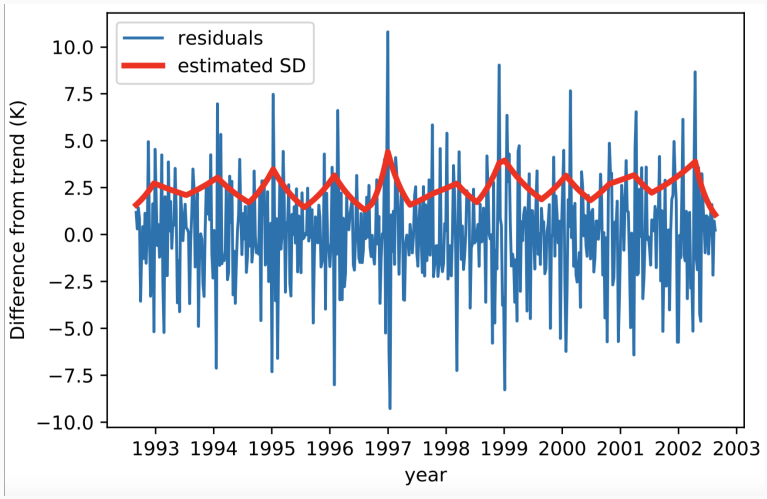
- Want a bound on $KL(\widehat{\theta} \,\|\, \theta_o)$
- Can use properties of exponential families to get "Basic inequality"

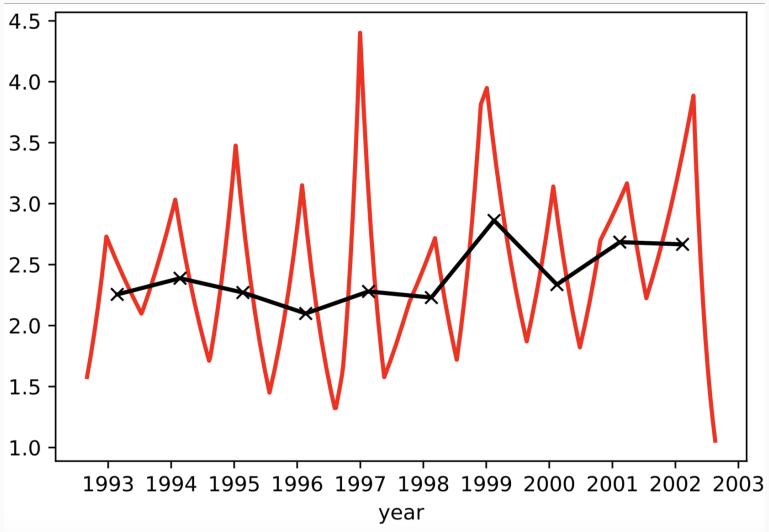
$$KL(\widehat{\theta} \,\|\, \theta_o) \leq (Y - A'(\theta_o))^\top (\theta_o - \widehat{\theta}) + \lambda \|D\theta_o\| - \lambda \left\|D\widehat{\theta}\right\|$$

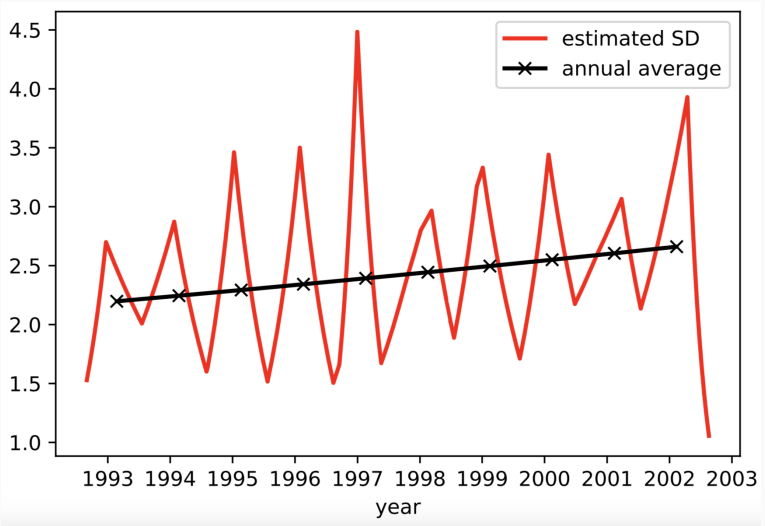
- First term is empirical process, second term controlled by λ
- $Y - A'(\theta_o)$ is mean zero, sub-exponential
- Play some games

Empirical results

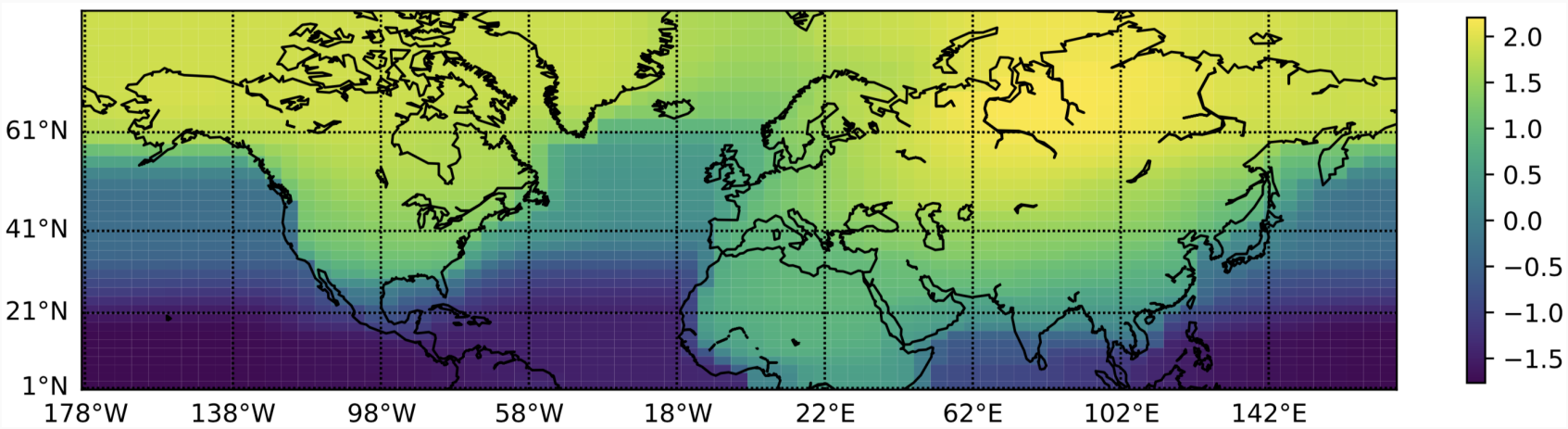
Bloomington, IN



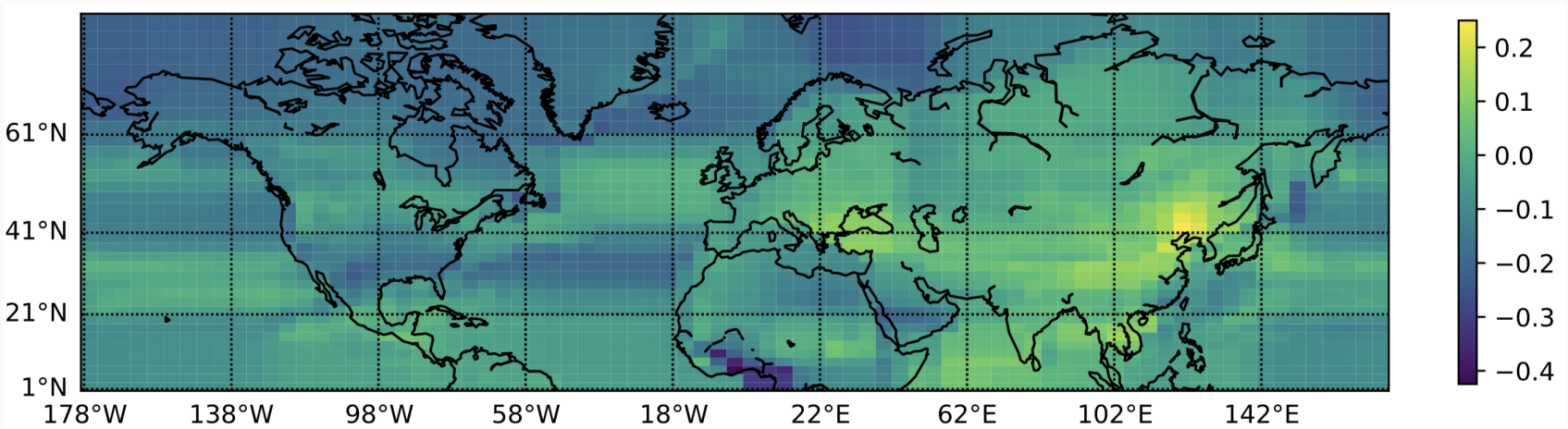




Average variance in base period



Change in average variance from 1961–2011



Conclusion

Wrapping up

- Generalized TF to exponential families
- Tailored algorithms for some big data
- Tuning parameters in this setting are challenging
- Lot's of missing details about the actual data
- Do we care about θ? $A'(\theta)$?

Collaborators and funding