Trend filtering in exponential families

Daniel J. McDonald
Indiana University, Bloomington
dajmcdon.github.io
4 March 2020
These are my cats
Number of vomits/day

dosage
- 2x/day
- 1x/day
- every other day
- 2x/week


# vomits / day
$y_i$ is the number of vomits on day $i$

Poisson distributed with time-varying parameter $\phi_i$

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

**Goal:** estimate $\phi$ from data, $\phi$ should be “smooth”.

Set $\theta_i = \log \phi_i$

$$\minimize_{\theta \in \mathbb{R}^n} \mathbf{1}^T \exp(\theta) - y^T \theta + \lambda \|D\theta\|_1$$

$D$ matrix encodes smoothness
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
• \( \lambda \) 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 $\lambda$ 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 $\lambda$ reasonably
3. Near-minimax theoretical guarantees

Motivated by a climate change study
Estimating the trend in cloud-top temperature volatility
The scientific consensus is that

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

Global warming → 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
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.
• Weather satellites aren’t made for this.
• More information in higher moments than in average?
Satellite data

Once collaborators do lots of processing...

- 52,000 time series
- daily records over $\sim 50$ years
- “trends” are local, nonlinear, not sinusoidal
Let $X_{ijt}$ be the observed temperature at time $t$ and location $(i, j)$.

Suppose $X_{ijt} \sim \text{Normal}(0, \sigma_{ijt}^2)$

(Follows sophisticated detrending)

Estimate $\sigma^2$, but it should be “smooth” relative to space and time.

Use a matrix $D + \text{penalty}$ to encode this smoothness.
Exponential families, standard examples

- Gaussian
- Gamma
- Beta
- Von Mises
- Poisson
- Binomial
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 $(Y, \theta)$ instead of $(X, \phi)$ is the “natural” parameterization
Trend filtering
Optimization problem

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

\[
\min_{\theta \in \Theta} 1^T A(\theta) - y^T \theta + \lambda \|D\theta\|_1
\]
Optimization problem

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

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

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

$$\min_{\mu \in \mathbb{R}^n} \frac{1}{2} \|x - \mu\|^2 + \lambda \|D\mu\|_1 = \min_{\theta \in \mathbb{R}^n} \frac{1}{2} \theta^T \theta - y^T \theta + \lambda \|D\theta\|_1$$
Optimization problem

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

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

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

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

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

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

$$\theta = -\frac{1}{2\sigma^2}, y = x^2, \text{ and } A(z) = -\frac{1}{2} \log(-z)$$
Smoothness and penalty order, $D$ matrices

Constant, $k=0$

Linear, $k=1$

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

- Estimated theta
- 1st derivative
- 2nd derivative

Dec '18
Jun '19
Dec '19
Dec '18
Jun '19
Dec '19
Dec '18
Jun '19
Dec '19
Relations to other (similar) methods

Locally adaptive regression splines

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

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

Smoothing splines

\[
\min_{f \in \mathcal{W}_{(k+1)/2}} \frac{1}{2n} \|y - f\|_2^2 + \lambda \int \chi \left( f^{(\frac{h+1}{2})}(t) \right)^2 dt
\]

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

see 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 $\rightarrow \text{df} = p + 1$

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

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

Smoothing splines have same degrees of freedom
Local adaptivity
Local adaptivity

trendfilter, df=50  spline, df=50  spline, df=90
Algorithms
Optimization problem

\[
\min_{\theta} 1^T A(\theta) - y^T \theta + \lambda \|D\theta\|_1
\]

Standard optimizer: Primal Dual Interior Point method

Alternatively: Alternating Direction Method of Multipliers

see Kim et al. (2009); Tibshirani (2014)
Alternating direction method of multipliers

Restate the problem

Original

\[ \min_x f(x) + g(x) \]

Equivalent

\[ \min_{x,z} f(x) + g(z) \]

s.t. \( x - z = 0 \)

Then, iterate the following:

\[ x \leftarrow \arg\min_x f(x) + \frac{\rho}{2} \| x - z + u \|_2^2 \]

\[ z \leftarrow \arg\min_z g(z) + \frac{\rho}{2} \| x - z + u \|_2^2 \]

\[ u \leftarrow u + x - z \]
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

\[
\min_{\theta} \quad \frac{1}{2} \theta^T \theta - y^T \theta + \lambda \|D\theta\|_1
\]

Equivalent

\[
\min_{\theta, \alpha} \quad \frac{1}{2} \theta^T \theta - y^T \theta + \lambda \|\alpha\|_1
\]

s.t. \quad D\theta - \alpha = 0

\[
\theta \leftarrow \arg\min_{\theta} \frac{1}{2} \theta^T \theta - y^T \theta + \frac{\rho}{2} \|\alpha - D\theta + u\|_2^2
\]

\[
\alpha \leftarrow \arg\min_{\alpha} \lambda \|\alpha\|_1 + \frac{\rho}{2} \|D\theta - \alpha + u\|_2^2
\]

\[
u \leftarrow u - D\theta + \alpha
\]
Decoupling example (Gaussian mean)

Original

\[
\min_\theta \frac{1}{2} \theta^T \theta - y^T \theta + \lambda \|D\theta\|_1
\]

Equivalent

\[
\begin{align*}
\min_{\theta, \alpha} & \quad \frac{1}{2} \theta^T \theta - y^T \theta + \lambda \|\alpha\|_1 \\
\text{s.t.} & \quad D\theta - \alpha = 0
\end{align*}
\]

\[\theta \leftarrow \text{matrix multiply}\]
\[\alpha \leftarrow \text{elementwise soft-threshold}\]
\[u \leftarrow \text{add vectors}\]
Decoupling example (Gaussian mean)

**Original**

\[
\min_{\theta} \quad \frac{1}{2} \theta^T \theta - y^T \theta + \lambda \|D\theta\|_1
\]

**Equivalent**

\[
\min_{\theta, \alpha} \quad \frac{1}{2} \theta^T \theta - y^T \theta + \lambda \|\alpha\|_1
\]

s.t. \quad D\theta - \alpha = 0

\[
\theta \leftarrow (I_n + \rho D^T D)^{-1} (y + \rho D^T (\alpha + u))
\]

\[
\alpha \leftarrow S_{\lambda/\rho} (D\theta + u)
\]

\[
u \leftarrow u - D\theta + \alpha
\]

\[
[S_a(b)]_k = \text{sign}(b_k)(|b_k| - a)_+
\]
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?

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

$D^TD$ occupies 8000 Petabytes, and you have to invert it.

Need custom algorithms/code
Requires very few iterations, but iterations cost $O(|\text{block}|/\text{three.osf})$ and can parallelize over blocks.
Consensus version

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

$x_g \leftarrow$ use PDIP on smaller blocks

$\theta \leftarrow$ average over groups

$u_g \leftarrow$ add vectors
Grid world

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

/two.osf/nine.osf
\( \theta_{ijt} \leftarrow \text{find a root} \\
\text{each line } \leftarrow \text{1D TF with the convex loss} \\
\text{dual variables } \leftarrow \text{add vectors}

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

We develop two new ADMM-type algorithms

Choice depends on computing architecture

Simulations: 4 sec vs 2 hours at 400 iterations

Smaller problems don’t need these details

Must repeat for many tuning parameters

see Khodadadi and McDonald (2019) for details
Tuning parameter selection
Unbiased risk estimation

$$\text{MSE}(\lambda) = \mathbb{E} \left[ \left\| \theta_0 - \hat{\theta}_\lambda(Y) \right\|_2^2 \right]$$

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

$$\text{MSE}(\lambda) = \mathbb{E} \left[ \left\| \theta_0 - \hat{\theta}_\lambda(Y) \right\|_2^2 \right] - n \sigma + \frac{1}{2} \sigma^2 \text{tr} \left( W \right), \quad \text{df} = \frac{1}{\sigma^2} \text{tr} \left( W \right)$$

e.g. Efron (1986)
Unbiased risk estimation

\[
\text{MSE}(\lambda) = \mathbb{E} \left[ \left\| \theta_0 - \hat{\theta}_\lambda(Y) \right\|_2^2 \right]
\]

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

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

e.g. Efron (1986)
Unbiased risk estimation

$$\text{MSE}(\lambda) = \mathbb{E} \left[ \left\| \theta_o - \hat{\theta}_\lambda(Y) \right\|^2 \right]$$

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

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

If $\hat{\theta}_\lambda(y) = W y$, then $\text{tr} \text{ Cov} \left( Y, \hat{\theta}_\lambda(Y) \right) = \sigma^2 \text{tr} (W)$

$$\overline{\text{MSE}}(\lambda) = \left\| Y - \hat{\theta}_\lambda(Y) \right\|^2_2 - n \sigma^2 + 2 \text{df}, \quad \text{df} := \frac{1}{\sigma^2} \text{tr}(W)$$

e.g. Efron (1986)
Stein (1981):

- Assume \( Y \sim \text{Normal}(\theta_0, \sigma^2 I_n) \)

+ \( \hat{\theta}_\lambda(Y) \) weakly differentiable
Extensions

Stein (1981):

- Assume $Y \sim \text{Normal}(\theta_0, \sigma^2 I_n)$

+ $\hat{\theta}_\lambda(Y)$ weakly differentiable

Eldar (2009):

+ Assume $Y \sim \text{ExpFam}(\theta_0)$, continuous (a.e.)

+ $\hat{\theta}_\lambda(Y)$ weakly differentiable
Extensions

Stein (1981):
- Assume $Y \sim \text{Normal}(\theta, \sigma^2 I_n)$
+ $\hat{\theta}_\lambda(Y)$ weakly differentiable

Eldar (2009):
+ Assume $Y \sim \text{ExpFam}(\theta)$, continuous (a.e.)
+ $\hat{\theta}_\lambda(Y)$ weakly differentiable

Both cases

1. Unbiased estimator of $\text{MSE}(\lambda)$
2. Need to know $\frac{\partial \hat{\theta}_\lambda}{\partial Y_i}(Y)$, the divergence
Extensions

Stein (1981):
- Assume $Y \sim \text{Normal}(\theta, \sigma^2 I_n)$
+ $\widehat{\theta}_\lambda(Y)$ weakly differentiable

Eldar (2009):
+ Assume $Y \sim \text{ExpFam}(\theta)$, continuous (a.e.)
+ $\widehat{\theta}_\lambda(Y)$ weakly differentiable

Both cases

1. Unbiased estimator of $\text{MSE}(\lambda)$
2. Need to know $\frac{\partial \widehat{\theta}_\lambda_i}{\partial Y_i}(Y)$, the divergence

Problems: (1) We don’t want the MSE. (2) We don’t know the divergence.
Estimating KL

Stein KL Estimator:

$$\widehat{KL}(\theta_0 \| \hat{\theta}_\lambda) = \left\langle \hat{\theta}_\lambda + \frac{h'(y)}{h(y)}, A'(\hat{\theta}_\lambda) \right\rangle + \left\langle A''(\hat{\theta}_\lambda), \frac{\partial \hat{\theta}_\lambda, i}{\partial y_i}(y) \right\rangle - \mathbf{1}^T A(\hat{\theta}_\lambda)$$

see Deledalle (2017)
Stein KL Estimator:

\[
\hat{KL}(\theta_0 \parallel \hat{\theta}_\lambda) = \left\langle \hat{\theta}_\lambda + \frac{h'(y)}{h(y)}, A'(\hat{\theta}_\lambda) \right\rangle + \left\langle A''(\hat{\theta}_\lambda), \frac{\partial \hat{\theta}_{\lambda,i}}{\partial y_i}(y) \right\rangle - 1^T A(\hat{\theta}_\lambda)
\]

with \( \mathbb{E} \left[ \hat{KL}(\theta_0 \parallel \hat{\theta}_\lambda) \right] = KL(\theta_0 \parallel \hat{\theta}_\lambda) - A(\theta_0) \).
Stein KL Estimator:

\[ \hat{KL}(\theta_o \| \hat{\theta}_\lambda) = \left< \hat{\theta}_\lambda + \frac{h'(y)}{h(y)}, \ A'(\hat{\theta}_\lambda) \right> + \left< A''(\hat{\theta}_\lambda), \ \frac{\partial \hat{\theta}_{\lambda,i}}{\partial y_i}(y) \right> - \mathbf{1}^T A(\hat{\theta}_\lambda) \]

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

Solves 1.

see Deledalle (2017)
Estimating KL

Stein KL Estimator:

\[
\hat{KL}(\theta_0 \parallel \hat{\theta}_\lambda) = \left\langle \hat{\theta}_\lambda + \frac{h'(y)}{h(y)}, A'(\hat{\theta}_\lambda) \right\rangle + \left\langle A''(\hat{\theta}_\lambda), \frac{\partial \hat{\theta}_{\lambda,i}}{\partial y_i}(y) \right\rangle - \mathbf{1}^T A(\hat{\theta}_\lambda)
\]

with \( \mathbb{E} \left[ \hat{KL}(\theta_0 \parallel \hat{\theta}_\lambda) \right] = KL(\theta_0 \parallel \hat{\theta}_\lambda) - A(\theta_0). \)

Solves 1.

Variance estimation:

\[
\hat{KL}(\theta_0 \parallel \hat{\theta}_\lambda) = \frac{1}{4} \left\langle y, \hat{\theta}_\lambda^{-1} \right\rangle + \left\langle \hat{\theta}_\lambda^{-2}, \frac{\partial \hat{\theta}_{\lambda,i}}{\partial y_i}(y) \right\rangle + \frac{1}{2} \mathbf{1}^T \log(-\hat{\theta}_\lambda) - \frac{1}{2}
\]

see Deledalle (2017)
Estimating KL

Stein KL Estimator:

\[
\hat{KL} \left( \theta_0 \parallel \hat{\theta}_\lambda \right) = \left\langle \hat{\theta}_\lambda + \frac{h'(y)}{h(y)}, A' \left( \hat{\theta}_\lambda \right) \right\rangle + \left\langle A''(\hat{\theta}_\lambda), \frac{\partial \hat{\theta}_{\lambda,i}}{\partial y_i}(y) \right\rangle - 1^T A(\hat{\theta}_\lambda)
\]

with \( \mathbb{E} \left[ \hat{KL} \left( \theta_0 \parallel \hat{\theta}_\lambda \right) \right] = KL \left( \theta_0 \parallel \hat{\theta}_\lambda \right) - A(\theta_0) \).

Solves 1.

Variance estimation:

\[
\hat{KL} \left( \theta_0 \parallel \hat{\theta}_\lambda \right) = \frac{1}{4} \left\langle y, \hat{\theta}_\lambda^{-1} \right\rangle + \left\langle \hat{\theta}_\lambda^{-2}, \frac{\partial \hat{\theta}_{\lambda,i}}{\partial y_i}(y) \right\rangle + \frac{1}{2} 1^T \log(-\hat{\theta}_\lambda) - \frac{1}{2}
\]

see Deledalle (2017)
Define $\Pi_D$, the projection onto the rows of $D$ with $D\hat{\theta} = 0$.

For trend filtering with exponential family loss:

$$\frac{\partial \hat{\theta}_{\lambda,i}(y)}{\partial y_i} = \left( \left( \Pi_D \text{diag} \left( A''(\hat{\theta}_\lambda) \right) \Pi_D \right)^\dagger \right)_{ii}$$
The divergence (our result)

Define $\Pi_D$, the projection onto the rows of $D$ with $D\hat{\theta} = 0$.

For trend filtering with exponential family loss:

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

Solves 2.
Define $\Pi_D$, the projection onto the rows of $D$ with $D\hat{\theta} = 0$.

For trend filtering with exponential family loss:

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

Solves 2.

Variance estimation: $A''(\theta) = \frac{1}{2\theta^2}$

$$\hat{KL} \left( \theta_\circ \parallel \hat{\theta}_\lambda \right) = -\frac{1}{2} + \sum_i \frac{y_i}{4\hat{\theta}_{\lambda,i}} + \frac{2 \left( (\Pi_D \text{diag} \left( \hat{\theta}_{\lambda, \lambda}^{-2} \right) \Pi_D) \right)^{\dagger}_{ii} + \log(-\hat{\theta}_{\lambda,i})}{2}$$
- Compare to Gaussian case: $\hat{d}_f = \text{tr}(\Pi_D)$ (Tibshirani and Taylor, 2012)

+ Measures the curvature correctly (compared to MSE)

+ No sample splitting, recomputing

+ Interpretable

+ Estimates the risk we control theoretically
Theory
1. $\lambda_n$ is large enough to control the empirical process
2. $\theta_0$ is $k$-times differentiable, and $\text{TV}(\theta_0^{(k)}) < C_n$
3. Observations on a $d$-dimensional regular grid
4. Ignore log factors which are myriad and ugly

Theorem:

$$\frac{1}{n} \text{KL} \left( \theta_0 \parallel \hat{\theta}_{\lambda_n} \right) = \begin{cases} O_p \left( \left( \frac{1}{n} \right)^{\frac{k+1}{d}} \right) & d \geq 2k + 2 \\ O_p \left( \left( \frac{1}{n} \right)^{\frac{2k+2}{2k+2+d}} \right) & d < 2k + 2 \end{cases}$$
Convergence result

1. $\lambda_n$ is large enough to control the empirical process
2. $\theta_0$ is $k$-times differentiable, and $\text{TV}(\theta_0^{(k)}) < C_n$
3. Observations on a $d$-dimensional regular grid
4. Ignore log factors which are myriad and ugly

Theorem:

$$\frac{1}{n} \text{KL} \left( \theta_0 \parallel \hat{\theta}_n \right) = \begin{cases} O_p \left( \left( \frac{1}{n} \right)^{\frac{k+1}{d}} \right) & d \geq 2k + 2 \\ O_p \left( \left( \frac{1}{n} \right)^{\frac{2k+2}{2k+2+d}} \right) & d < 2k + 2 \end{cases}$$
Notes on our theorem

\[
\frac{1}{n} \text{KL} \left( \theta_0 \parallel \hat{\theta}_{\lambda_n} \right) = \begin{cases} 
O_p \left( \left( \frac{1}{n} \right)^{\frac{k+1}{d}} \right) & d \geq 2k + 2 \\
O_p \left( \left( \frac{1}{n} \right)^{\frac{2k+2}{2k+2+d}} \right) & d < 2k + 2
\end{cases}
\]

- 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”

\[
KL \left( \theta_0 \| \hat{\theta} \right) \leq (Y - A'(\theta_0))^\top(\theta_0 - \hat{\theta}) + \lambda \|D\theta_0\| - \lambda \|D\hat{\theta}\|
\]

• 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”

\[ KL \left( \theta_0 \parallel \hat{\theta} \right) \leq (Y - A'(\theta_o))^\top (\theta_0 - \hat{\theta}) + \lambda \| D\theta_0 \| - \lambda \| D\hat{\theta} \| \]

• First term is empirical process, second term controlled by $\lambda$

• $Y - A'(\theta_o)$ is mean zero, sub-exponential

• Play some games

\[ \ldots 15 \text{ pages of } \LaTeX \ldots \]
Empirical results
Toronto temperature

°C

1960 1980 2000
Toronto temperature
Change in estimated SD (1960s vs 2000s)

Summer

Winter

°C

-2

-1

0

1

2
Change in mean temperature (1960s vs 2000s)

- Temperature change for summer:
  - -12°C
  - -4°C
  - -1°C
  - 0°C
  - 1°C
  - 4°C
  - 12°C

- Temperature change for winter:
  - -12°C
  - -4°C
  - -1°C
  - 0°C
  - 1°C
  - 4°C
  - 12°C
Observed temperatures in Toronto (1960s vs 2000s)
Conclusion
We generalized TF to exponential families

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

Future work

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

Computational choices impact scientific conclusions

These choices can take many forms:

- selecting tuning parameters
- different optimization algorithms return different solutions
- how long do we run our MCMC (and which kind do we use)

Statistical theory often neglects these choices:

- LASSO works with oracle tuning parameter
- We have the posterior if our MCMC runs forever
- EM gives us a global solution
Applications demand techniques that couple

1. computational considerations
2. statistical regularization
Applications demand techniques that couple

1. computational considerations
2. statistical regularization

Therefore, two important questions must be addressed:

1. How does the algorithm impact the science?
2. How do we select tuning parameters when computations are at a premium?
My research program seeks...

1. to enable application through reasoned tuning parameter selection; (Homrighausen and McDonald, 2013, 2014, 2017, 2018)
My research program seeks...

1. to enable application through reasoned tuning parameter selection; (Hmrighausen and McDonald, 2013, 2014, 2017, 2018)

2. to deepen the theoretical understanding of approximate algorithms; (Ding and McDonald, 2017, 2019; Hmrighausen and McDonald, 2016, 2019)
My research program seeks...

1. to enable application through reasoned tuning parameter selection; (Homrighausen and McDonald, 2013, 2014, 2017, 2018)

2. to deepen the theoretical understanding of approximate algorithms; (Ding and McDonald, 2017, 2019; Homrighausen and McDonald, 2016, 2019)

3. to develop approximation and tuning parameter selection techniques for dependent data; (McDonald, 2019; McDonald and Shalizi, 2019a,b; McDonald et al., 2011, 2015)
1. to enable application through reasoned tuning parameter selection; (Homrighausen and McDonald, 2013, 2014, 2017, 2018)

2. to deepen the theoretical understanding of approximate algorithms; (Ding and McDonald, 2017, 2019; Homrighausen and McDonald, 2016, 2019)

3. to develop approximation and tuning parameter selection techniques for dependent data; (McDonald, 2019; McDonald and Shalizi, 2019a,b; McDonald et al., 2011, 2015)

4. to characterize the effects of algorithmic or other approximations in nonparametrics; (McDonald, 2017; McDonald et al., 2017, 2019a)
1. to enable application through reasoned tuning parameter selection; (Homrighausen and McDonald, 2013, 2014, 2017, 2018)

2. to deepen the theoretical understanding of approximate algorithms; (Ding and McDonald, 2017, 2019; Homrighausen and McDonald, 2016, 2019)

3. to develop approximation and tuning parameter selection techniques for dependent data; (McDonald, 2019; McDonald and Shalizi, 2019a,b; McDonald et al., 2011, 2015)

4. to characterize the effects of algorithmic or other approximations in nonparametrics; (McDonald, 2017; McDonald et al., 2017, 2019a)

5. to apply the proposed tools to meaningful applications. (Ding and McDonald, 2017, 2019; Khodadadi and McDonald, 2019; McDonald and Shalizi, 2019a; McDonald et al., 2019b)
Research overview

How do we select tuning parameters when computations are at a premium?

How does the algorithm impact the science?
How do we select tuning parameters when computations are at a premium?

How does the algorithm impact the science?

My research program seeks to demonstrate

1. How to select tuning parameters in various contexts.
2. How algorithms can enable scientific conclusions.
3. How we can use approximate algorithms to improve some inferential procedures.
Collaborators and funding
Appendix
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
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 $\theta$

Banded for TF

So 2 and 3 are solved in linear time.
Detailed PDIP

**Primal**

\[
\min_{\theta} \quad f(\theta) + \lambda \|D\theta\|_1
\]

**Dual**

\[
\min_{v} \quad f^*(-D^T v)
\]

s.t. \( \|v\|_\infty \leq \lambda \)

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

Perturbed KKT conditions \((w > 0) \implies r_w(v, \mu_1, \mu_2) := \begin{bmatrix} \nabla f^*(-D^T v) + D(v - \lambda 1)^T \mu_1 - D(v + \lambda 1)^T \mu_2 \\ -\mu_1(v - \lambda 1) + \mu_2(v + \lambda 1) - w^{-1}1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \)

- As \( w \rightarrow \infty \), 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

\[
\min_{f \in \mathcal{F}_k} \frac{1}{2n} \| y - f \|^2 + \lambda \text{TV}(f^{(k)})
\]

- \( \mathcal{F}_k = \{ f : [0, 1] \to \mathbb{R}, f^{(k)} \text{ exists a.e.}, \text{TV}(f^{(k)}) < \infty \} \)

- 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

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

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

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

- \( W_{(k+1)/2} = \{ f : [0, 1] \to \mathbb{R}, f^{(k)} \text{ exists}, \int_{\mathcal{X}} \left( f^{(k+1)/2}(t) \right)^2 dt < \infty \} \)

- Solution is a \( k^{th} \)-degree spline (Wahba, 1990)

- \( k \) needs to be odd

- One way to solve:

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

- \( K \) is banded, so solution requires \( O(n) \) computations.
What our data look like

cylindrical projection
\[
\min_{x_g = \theta} \forall g \in G \sum_{g \in G} -\ell(x_g) + \lambda \|D_g \cdot x_g\|_1
\]

\[
x_g \leftarrow \arg\min_{x_g} -\ell(x_g) + \lambda \|D_g \cdot x_g\|_1
\]

\[
+ u^T(x_g - \theta) + \frac{\rho}{2} \|x_g - \theta\|_2^2
\]

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

\[
u_g \leftarrow u_g + \rho(x_g - \theta)
\]
\[
\min_{\theta=a=b=c} \sum_{ijt} -\ell(\theta_{ijt}) + \lambda \sum_{it} \| Da_{i,t} \|_1 \\
+ \lambda \sum_{jt} \| Db_{j,t} \|_1 + \lambda \sum_{ij} \| Dc_{ij} \|_1
\]

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

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

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

\[
k^{(1)}, k^{(2)} \leftarrow \text{simple linear functions of } a, b, c, u, v, w
\]
Stein’s unbiased risk estimator

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

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

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

(Stein, 1981)
Generalized SURE for continuous exp fam

- If \( p_\theta(y) = h(y) \exp(\theta^T y - 1^T A(\theta)) \)
- And \( h(\cdot) \) is weakly differentiable

\[
\mathbb{E} \left[ \theta_0^T \hat{\theta}_\lambda(Y) \right] = -\mathbb{E} \left[ \left( \frac{h'(Y)}{h(Y)}, \hat{\theta}_\lambda(Y) \right) + \sum_i \left( \frac{\partial \hat{\theta}_\lambda,i}{\partial Y_i}(Y) \right) \right]
\]

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

\[
\left\| \hat{\theta}_\lambda \right\|_2^2 + 2 \left( \frac{h'(y)}{h(y)} \right)^T \hat{\theta}_\lambda + 2 \sum_i \left( \frac{\partial \hat{\theta}_\lambda,i}{\partial y_i}(y) \right) + \text{tr} \left( h''(y) \right) \frac{h(y)}{h(y)}
\]

(Eldar, 2009)
The Divergence

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

For TF for Gaussian mean:

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

(Tibshirani and Taylor, 2012)
The Divergence

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

For TF for Gaussian mean:

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

Count the pieces + $k + 1$

(Tibshirani and Taylor, 2012)
Which classes and canonical scaling

- $D$ is such that it smooths over axis parallel lines in the grid
- Define $\mathcal{K}^k_d(C_n) = \{\theta : \|D\theta\|_1 < C_n\}$
- Define $\mathcal{H}^{k+1}_d(L)$ to be the Hölder class containing discretized Hölder smooth-functions with $k$ derivatives
- Can show that $\mathcal{H}^{k+1}_d(L) \subset \mathcal{K}^k_d(cn^{1-(k+1)/d})$
- This gives the lower bound.
- Linear smoothers can’t achieve this rate (Donoho and Johnstone, 1998)
Like LASSO other $\ell_1$-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)
1. to enable application through reasoned tuning parameter selection;

- SURE for logistic regression. McDonald and Tibshirani. (in progress)
- Approximate Rademacher Complexities. McDonald. (in progress)
1. to enable application through reasoned tuning parameter selection;


- SURE for logistic regression. McDonald and Tibshirani. (in progress)

- Approximate Rademacher Complexities. McDonald. (in progress)
CV “works” for lasso

Under strong conditions

\[ \mathbb{E} \left[ \left( Y_0 - X_0^T \hat{\beta}_\lambda \right)^2 \right] = O_p \left( \frac{s \log(p) \log(n)}{n} \right) \]

Under weak conditions

\[ \mathbb{E} \left[ \left( Y_0 - X_0^T \hat{\beta}_t \right)^2 \right] - \mathbb{E} \left[ \left( Y_0 - X_0^T \beta_t \right)^2 \right] = o(1) \]

for \( t_n = o \left( \left( \frac{n}{\log(p) \log(n)} \right)^{1/4} \right), \| \beta \|_1 \leq t_n. \)
CV “works” for lasso

Under strong conditions

\[ \mathbb{E} \left[ \left( Y - X_0^T \hat{\beta}_\lambda \right)^2 \right] = O_P \left( \frac{s \log(p) \log(n)}{n} \right) \]

Under weak conditions

\[ \mathbb{E} \left[ \left( Y - X_0^T \hat{\beta}_t \right)^2 \right] - \mathbb{E} \left[ \left( Y - X_0^T \beta_t \right)^2 \right] = o(1) \]

for \( t_n = o \left( \left( \frac{n}{\log(p) \log(n)} \right)^{1/4} \right) \), \( \| \beta \|_1 \leq t_n \).

CV "costs" \( \log(n) \).
My research program seeks...

2. to deepen the theoretical understanding of approximate algorithms;


- Predicting phenotypes from microarrays using amplified, initially marginal, eigenvector regression. Ding and McDonald. *Bioinformatics*. (2017)

- Sufficient principal component regression. Ding and McDonald. (submitted)

- Semi-supervised learning in high dimensions with structured manifolds. Ding. (2020, PhD thesis)

- Compression improves estimation under model misspecification. McDonald. (in progress)
2. to deepen the theoretical understanding of approximate algorithms;


- Predicting phenotypes from microarrays using amplified, initially marginal, eigenvector regression. Ding and McDonald. *Bioinformatics*. (2017)

- Sufficient principal component regression. Ding and McDonald. (submitted)

- Semi-supervised learning in high dimensions with structured manifolds. Ding. (2020, PhD thesis)

- Compression improves estimation under model misspecification. McDonald. (in progress)
2. to deepen the theoretical understanding of approximate algorithms;


- Predicting phenotypes from microarrays using amplified, initially marginal, eigenvector regression. Ding and McDonald. *Bioinformatics*. (2017)

- Sufficient principal component regression. Ding and McDonald. (submitted)

- Semi-supervised learning in high dimensions with structured manifolds. Ding. (2020, PhD thesis)

- Compression improves estimation under model misspecification. McDonald. (in progress)
Suppose $y_i = x_i^\top \beta^* + \epsilon_i$

Previous work:

- Assume that $\text{Cov}(y, X_j) = 0 \Rightarrow \beta_j^* = 0$.
- Algorithm: 1. screen by covariance, 2. perform PCR

Our work:

- Note that $\left\| \nu \left( \mathbb{E} \left[ X^\top X \right] \right) \nu \right\|^2 = 0 \Rightarrow \beta_j^* = 0$.
- Algorithm: 1. Perform regularized PCR

(Bair and Tibshirani, 2004; Bair et al., 2006; Paul et al., 2008; Tay et al., 2018)
Suppose \( y_i = x_i^T \beta^* + \epsilon_i \)

Previous work:

- Assume that Cov\((y, X_j) = 0 \Rightarrow \beta^*_j = 0.\)
- Algorithm: 1. screen by covariance, 2. perform PCR

Our work:

- Note that \( \| v \left( \mathbb{E} \left[ X^T X \right] \right) \|_2 = 0 \Rightarrow \beta^*_j = 0. \)
- Algorithm: 1. Perform regularized PCR

Intuition:

\[
\beta^* = \mathbb{E} \left[ X^T X \right]^{-1} \mathbb{E} \left[ X^T y \right] = VD^{-2} V^T VDU^T y = VD^{-1} U^T y
\]

(Bair and Tibshirani, 2004; Bair et al., 2006; Paul et al., 2008; Tay et al., 2018)
Sufficient PCR

MSE

# Features Selected

Precision

Recall

SuffPCR

Oracle

Lasso

Ridge

ElasticNet

SPC

ISPCA

FPS

 scatter plots showing comparisons between different methods in terms of MSE, number of features selected, precision, and recall.
Theorem

Assume many conditions, $s := |\beta_*|$, $\text{supp}(v) := \{j : v_j \neq 0\}$,

$$\left\| \mathbf{x} \left( \hat{\beta} - \beta_* \right) \right\|_2 = O_P \left( \sigma \sqrt{\frac{(s^2 + d) \log p}{n}} \right),$$

and

$$\left| \text{supp}(\hat{\beta}) \Delta \text{supp}(\beta_*) \right| = O_P \left( \sigma \frac{s^2 \log p}{n} \right).$$
This methodology uses two insights from earlier work (Hormighausen and McDonald, 2016, 2019)

1. Random projection works well when it gets the columns that have the most information.

2. SVD is computationally expensive. ADMM steps can be approximate under certain conditions.
3. to develop approximation algorithms for dependent data;

- Sparse additive state-space models. McDonald and Shalizi. (in progress)
- Empirical macroeconomics and DSGE modeling in statistical perspective. McDonald and Shalizi. (in progress)
- Rademacher complexity of stationary sequences. McDonald and Shalizi. (submitted)
- Approximate Kalman Filtering. McDonald (in progress)
My research program seeks...

3. to develop approximation algorithms for dependent data;

- Sparse additive state-space models. McDonald and Shalizi. (in progress)
- Empirical macroeconomics and DSGE modeling in statistical perspective. McDonald and Shalizi. (in progress)
- Rademacher complexity of stationary sequences. McDonald and Shalizi. (submitted)
- Approximate Kalman Filtering. McDonald (in progress)
3. to develop approximation algorithms for dependent data;

- Sparse additive state-space models. McDonald and Shalizi. (in progress)
- Empirical macroeconomics and DSGE modeling in statistical perspective. McDonald and Shalizi. (in progress)
- Rademacher complexity of stationary sequences. McDonald and Shalizi. (submitted)
- Approximate Kalman Filtering. McDonald (in progress)
Econ forecasting models don’t know “output” from “interest”

38% of permutations improved

change in prediction MSE
Economic forecasting models will never learn

<table>
<thead>
<tr>
<th>Investment</th>
<th>Output</th>
<th>Wages</th>
<th>Inflation</th>
<th>Interest Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>250</td>
<td>500</td>
<td>750</td>
<td>1000</td>
<td>250</td>
</tr>
<tr>
<td>500</td>
<td>250</td>
<td>500</td>
<td>750</td>
<td>500</td>
</tr>
<tr>
<td>750</td>
<td>750</td>
<td>750</td>
<td>750</td>
<td>750</td>
</tr>
<tr>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
</tr>
</tbody>
</table>

Change in prediction error relative to the truth

Number of training points

change in prediction error relative to the truth

250 500 750 1000 250 500 750 1000 250 500 750 ...
4. to characterize the effects of algorithmic or other approximations in nonparametrics;

- Exponential family trend filtering on grids. McDonald, Sharpnack, Bassett, and Sandhanala. (in progress)
- Minimax density estimation for growing dimension. McDonald. AISTATS. (2017)
- Minimax non-parametric regression with interactions. McDonald and Kolar. (in progress)
My research program seeks...

4. to characterize the effects of algorithmic or other approximations in nonparametrics;

• Exponential family trend filtering on grids. McDonald, Sharpnack, Bassett, and Sandhanala. (in progress)
• Minimax density estimation for growing dimension. McDonald. AISTATS. (2017)
• Nonparametric risk bounds for time-series forecasting. McDonald, Shalizi, and Shervish. JMLR. (2017)
• Minimax non-parametric regression with interactions. McDonald and Kolar. (in progress)
4. to characterize the effects of algorithmic or other approximations in nonparametrics;

- Exponential family trend filtering on grids. McDonald, Sharpnack, Bassett, and Sandhanala. (in progress)
- Minimax density estimation for growing dimension. McDonald. AISTATS. (2017)
- Minimax non-parametric regression with interactions. McDonald and Kolar. (in progress)
Suppose your data is supported on a low-dimensional manifold.

You don’t know the dimension, start small and increase as you collect more data.

No theory saying how to increase the dimension

Examples:

• PCA + density estimation, what $d$ to use?
• How many brain regions can we estimate a density over?
If \( p \geq 2, \exists 0 < a \leq A < \infty \) independent of \( d, n \) such that

\[
a \left( \frac{d^d}{n^\beta} \right)^{\frac{1}{2\beta + d}} \leq \inf_{\hat{f}} \sup_{f \in \mathcal{N}} \mathbb{E} \left[ \| \hat{f} - f \|_p \right] \leq \sup_{f \in \mathcal{N}} \mathbb{E} \left[ \| \hat{f}_h - f \|_p \right] \leq A \left( \frac{d^d}{n^\beta} \right)^{\frac{1}{2\beta + d}}.
\]

Consistency requires

\[
d = o \left( \frac{\beta \log n}{W(\beta \log n)} \right)
\]
If $p \geq 2$, there exists $a < A < \infty$ independent of $d, n$ such that

$$a \left( \frac{d^d}{n^\beta} \right)^{\frac{1}{2\beta + d}} \leq \inf_{\tilde{f}} \sup_{f \in \mathcal{N}} \mathbb{E} \left[ \left\| \tilde{f} - f \right\|_p \right] \leq \sup_{f \in \mathcal{N}} \mathbb{E} \left[ \left\| \hat{f}_h - f \right\|_p \right] \leq A \left( \frac{d^d}{n^\beta} \right)^{\frac{1}{2\beta + d}}.$$

Consistency requires

$$d = o \left( \frac{\beta \log n}{W(\beta \log n)} \right)$$
My research program seeks...

5. to apply the proposed tools to meaningful applications.

- Empirical macroeconomics and DSGE modeling in statistical perspective. McDonald and Shalizi. (in progress)
- Exponential family trend filtering on grids. McDonald, Sharpnack, Bassett, and Sandhanala. (in progress)
- Sparse facicle estimation from diffusion tensor imaging. McDonald, Cohen, …, Pestilli. (in progress)
- A switching model for vocal performances. Granger, McDonald, and Raphael. (in progress)
- Angular lasso for genetic clock time prediction. McDonald and Liu. (in progress)
5. to apply the proposed tools to meaningful applications.

- **Markov-switching state space models for uncovering musical interpretation.** McDonald, McBride, Gu, and Raphael. (submitted)
- **Empirical macroeconomics and DSGE modeling in statistical perspective.** McDonald and Shalizi. (in progress)
- **Cloud temperature time series analysis using state space approach.** Wang. (2017, MS thesis)
- **Exponential family trend filtering on grids.** McDonald, Sharpnack, Bassett, and Sandhanala. (in progress)
- **Sparse facicle estimation from diffusion tensor imaging.** McDonald, Cohen, …, Pestilli. (in progress)
- **A switching model for vocal performances.** Granger, McDonald, and Raphael. (in progress)
- **Angular lasso for genetic clock time prediction.** McDonald and Liu. (in progress)
5. to apply the proposed tools to meaningful applications.

- Empirical macroeconomics and DSGE modeling in statistical perspective. McDonald and Shalizi. (in progress)
- Exponential family trend filtering on grids. McDonald, Sharpnack, Bassett, and Sandhanala. (in progress)
- Sparse facicle estimation from diffusion tensor imaging. McDonald, Cohen, …, Pestilli. (in progress)
- A switching model for vocal performances. Granger, McDonald, and Raphael. (in progress)
- Angular lasso for genetic clock time prediction. McDonald and Liu. (in progress)
Clustering Chopin’s Mazurka with learned interpretations

Shebanova 2002
Wasowski 1980
Luisada 1991
Milkina 1970

Measure
Tempo
Constant
decelerate
accelerate
stress

/shebanova2002/
wasaki1980/
luisada1991/
milkina1970/


