Nonparametric density estimation on compact intervals with Bernstein polygrams

Jonas Moss

May 21, 2017

Getting into polygrams

Histograms and how to generalize them

We need Bernstein densities!

The goal: Bernstein polygrams

Two examples: Simulated and real

A bimodal mixture of Betas

A classic: Buffalo snowfall data (1910 - 1972)

The Power to Shape Your Density

Shape up with Bernstein polygrams

Generate uncommon shapes

The End

Histograms are mixtures of disjoint uniforms

- **Histograms**: Densities h like

$$h(x) = \sum_{i=1}^{K} \frac{w_i}{s_i - s_{i-1}} 1_{[s_{i-1}, s_i)}(x),$$

where $\sum w_i = 1$ and $s_0 < s_1 < \ldots < s_K$.

Histograms are mixtures of disjoint uniforms

- **Histograms**: Densities h like

$$h(x) = \sum_{i=1}^{K} \frac{w_i}{s_i - s_{i-1}} 1_{[s_{i-1}, s_i)}(x),$$

where $\sum w_i = 1$ and $s_0 < s_1 < \ldots < s_K$.

- A mixture of *disjoint uniform distributions*.

Histograms are mixtures of disjoint uniforms

- **Histograms**: Densities h like

$$h(x) = \sum_{i=1}^{K} \frac{w_i}{s_i - s_{i-1}} 1_{[s_{i-1}, s_i)}(x),$$

where $\sum w_i = 1$ and $s_0 < s_1 < \ldots < s_K$.

- A mixture of *disjoint uniform distributions*.
- Estimated by *maxiumum likehlihood* or *least squares* — same result either way: Estimated weights are $\widehat{w}_i = \#\mathrm{bin_i}/n$.

Must histograms be estimated by maximum likelihood?

- **Least squares**: $h(x) = \mathrm{argmin}_{h \in \mathcal{H}} \int (f(x) - h(x))^2 \, dx.$

Must histograms be estimated by maximum likelihood?

- **Least squares**: $h(x) = \mathrm{argmin}_{h \in \mathcal{H}} \int (f(x) - h(x))^2 \, dx$.
- Can be estimated empirically with

$$\mathrm{argmin}_{h \in \mathcal{H}} \left(-2 \frac{1}{n} \sum_{i=1}^{n} h(X_i) + \int h^2(x) \, dx \right)$$

Must histograms be estimated by maximum likelihood?

- **Least squares**: $h(x) = \mathrm{argmin}_{h \in \mathcal{H}} \int (f(x) - h(x))^2 \, dx$.
- Can be estimated empirically with

$$\mathrm{argmin}_{h \in \mathcal{H}} \left(-2 \frac{1}{n} \sum_{i=1}^{n} h(X_i) + \int h^2(x) \, dx \right)$$

- Histograms are agnostic to the choice of estimator — estimated weights are still $\widehat{w}_i = \#\mathrm{bin_i}/n$.

Must histograms be estimated by maximum likelihood?

- **Least squares**: $h(x) = \operatorname{argmin}_{h\in\mathcal{H}} \int (f(x) - h(x))^2\, dx$.
- Can be estimated empirically with

$$\operatorname{argmin}_{h\in\mathcal{H}} \left(-2\frac{1}{n}\sum_{i=1}^{n} h(X_i) + \int h^2(x)\, dx \right)$$

- Histograms are agnostic to the choice of estimator — estimated weights are still $\widehat{w}_i = \#\mathrm{bin_i}/n$.
- Makes sense to use! E.g. the Grenander estimator, convex densities, robust estimation.

What about disjoint mixtures of polynomials?

- For parameters θ_i and parametric density f on $[0,1]$ a generalized histogram based on f is

$$\sum_{k=1}^{K} \frac{w_i}{s_k - s_{k-1}} f\left(\frac{x - s_{k-1}}{s_k - s_{k-1}} \mid \theta_i\right) 1_{[s_{k-1}, s_k)}(x)\,.$$

What about disjoint mixtures of polynomials?

- For parameters θ_i and parametric density f on $[0,1]$ a generalized histogram based on f is

$$\sum_{k=1}^{K} \frac{w_i}{s_k - s_{k-1}} f\left(\frac{x - s_{k-1}}{s_k - s_{k-1}} \mid \theta_i\right) 1_{[s_{k-1}, s_k)}(x) .$$

- Natural choice for fs are *polynomial densities*.

What about disjoint mixtures of polynomials?

- For parameters θ_i and parametric density f on $[0,1]$ a generalized histogram based on f is

$$\sum_{k=1}^{K} \frac{w_i}{s_k - s_{k-1}} f\left(\frac{x - s_{k-1}}{s_k - s_{k-1}} \mid \theta_i\right) 1_{[s_{k-1},s_k)}(x) .$$

- Natural choice for fs are *polynomial densities*.
- **Problem:** Polynomial densities of form $\sum_{i=0}^{m} \beta_i x^i$ have complex constraints on β (involving m-dimensional ellipsoids).

What about disjoint mixtures of polynomials?

- For parameters θ_i and parametric density f on $[0,1]$ a generalized histogram based on f is

$$\sum_{k=1}^{K} \frac{w_i}{s_k - s_{k-1}} f\left(\frac{x - s_{k-1}}{s_k - s_{k-1}} \mid \theta_i\right) 1_{[s_{k-1},s_k)}(x).$$

- Natural choice for fs are *polynomial densities*.
- **Problem:** Polynomial densities of form $\sum_{i=0}^{m} \beta_i x^i$ have complex constraints on β (involving m-dimensional ellipsoids).
- **Solution 1**: Use families of polynomials with good properties. *Legendre* or *Bernstein*.

What about disjoint mixtures of polynomials?

- For parameters θ_i and parametric density f on $[0,1]$ a generalized histogram based on f is

$$\sum_{k=1}^{K} \frac{w_i}{s_k - s_{k-1}} f\left(\frac{x - s_{k-1}}{s_k - s_{k-1}} \mid \theta_i\right) 1_{[s_{k-1}, s_k)}(x) .$$

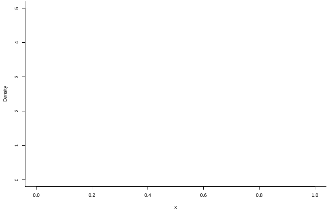
- Natural choice for fs are *polynomial densities*.
- **Problem:** Polynomial densities of form $\sum_{i=0}^{m} \beta_i x^i$ have complex constraints on β (involving m-dimensional ellipsoids).
- **Solution 1**: Use families of polynomials with good properties. *Legendre* or *Bernstein*.
- **Solution 2**: Use logsplines instead! (Cooperberg, Stone, 1990)

What is a Bernstein basis density?

A *Bernstein basis density* is

$$b_{\nu,m}(x) = (m+1)\binom{m}{\nu} x^{\nu}(1-x)^{m-\nu},$$

where $m, \nu \in \mathbb{N}$, $\nu \leq m$.

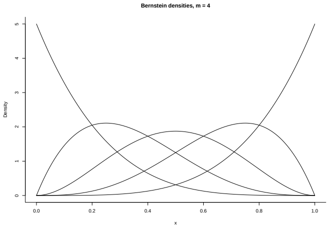


What is a Bernstein basis density?

A *Bernstein basis density* is

$$b_{\nu,m}(x) = (m+1)\binom{m}{\nu} x^{\nu}(1-x)^{m-\nu},$$

where $m, \nu \in \mathbb{N}$, $\nu \leq m$.



Some properties of Bernstein basis densities

1. **Span:** $\{b_{\nu,m}\}_{\nu=0}^{m}$ is a basis for the space of mth degree polynomials

Some properties of Bernstein basis densities

1. **Span:** $\{b_{\nu,m}\}_{\nu=0}^{m}$ is a basis for the space of mth degree polynomials
2. **Positive:** $\min b_{\nu,m}(x) = 0$; for all pairs (ν, m) and $x \in [0, 1]$

Some properties of Bernstein basis densities

1. **Span:** $\{b_{\nu,m}\}_{\nu=0}^{m}$ is a basis for the space of mth degree polynomials
2. **Positive:** $\min b_{\nu,m}(x) = 0$; for all pairs (ν, m) and $x \in [0, 1]$
3. **True densities:** $\int_0^1 b_{\nu,m}(x)\, dx = 1$ for all pairs (ν, m)

Some properties of Bernstein basis densities

1. **Span:** $\{b_{\nu,m}\}_{\nu=0}^{m}$ is a basis for the space of mth degree polynomials
2. **Positive:** $\min b_{\nu,m}(x) = 0$; for all pairs (ν, m) and $x \in [0, 1]$
3. **True densities:** $\int_0^1 b_{\nu,m}(x)\, dx = 1$ for all pairs (ν, m)
4. **Bounded maxima:** $\max b_{\nu,m}(x) \leq (m + 1)$ and $\max b_{m,m}(x) = \max b_{0,m}(x) = (m + 1)$

Some properties of Bernstein basis densities

1. **Span:** $\{b_{\nu,m}\}_{\nu=0}^{m}$ is a basis for the space of mth degree polynomials
2. **Positive:** $\min b_{\nu,m}(x) = 0$; for all pairs (ν, m) and $x \in [0,1]$
3. **True densities:** $\int_0^1 b_{\nu,m}(x)\,dx = 1$ for all pairs (ν, m)
4. **Bounded maxima:** $\max b_{\nu,m}(x) \leq (m+1)$ and $\max b_{m,m}(x) = \max b_{0,m}(x) = (m+1)$
5. **Beta distribution:** $b_{\nu,m}(x) = f_{\text{Beta}}(x \mid \nu+1, m-\nu+1)$

Some properties of Bernstein basis densities

1. **Span:** $\{b_{\nu,m}\}_{\nu=0}^{m}$ is a basis for the space of mth degree polynomials
2. **Positive:** $\min b_{\nu,m}(x) = 0$; for all pairs (ν, m) and $x \in [0,1]$
3. **True densities:** $\int_0^1 b_{\nu,m}(x)\, dx = 1$ for all pairs (ν, m)
4. **Bounded maxima:** $\max b_{\nu,m}(x) \leq (m+1)$ and $\max b_{m,m}(x) = \max b_{0,m}(x) = (m+1)$
5. **Beta distribution:** $b_{\nu,m}(x) = f_{\mathrm{Beta}}(x \mid \nu+1, m-\nu+1)$
6. **Pliable:** Rather simple expressions for derivatives, etc.

Bernstein densities are flexible densities!

1. For $m \in \mathbb{N}^+$ and λ in the unit m-simplex, the function

$$h(x \mid \lambda) = \sum_{\nu=0}^{m} \lambda_i b_{\nu,m}(x)$$

   is a polynomial density. A *convex combination* of Bernstein basis densities.

Bernstein densities are flexible densities!

1. For $m \in \mathbb{N}^+$ and λ in the unit m-simplex, the function

$$h(x \mid \lambda) = \sum_{\nu=0}^{m} \lambda_i b_{\nu,m}(x)$$

   is a polynomial density. A *convex combination* of Bernstein basis densities.
2. **Note:** Does *not* cover all polynomial densities of degree m. Still they are very flexible.

Bernstein densities have been used before.

1. **Maximum likelihood:** Studied by Guan (2016)

Bernstein densities have been used before.

1. **Maximum likelihood:** Studied by Guan (2016)
2. **Classical method:** Put $\lambda_\nu = F\left(\frac{\nu+1}{m+1}\right) - F\left(\frac{\nu}{m+1}\right)$ for a distribution function F. Motivated by Bernstein's proof of the Weierstraß approximation theorem. (Vitale, 1975)

Bernstein densities have been used before.

1. **Maximum likelihood:** Studied by Guan (2016)
2. **Classical method:** Put $\lambda_\nu = F\left(\frac{\nu+1}{m+1}\right) - F\left(\frac{\nu}{m+1}\right)$ for a distribution function F. Motivated by Bernstein's proof of the Weierstraß approximation theorem. (Vitale, 1975)
3. **Non-parametric Bayes:** Petrone (1999) allowed $F \sim \mathrm{DP}(F_0, \alpha)$ and $m \sim p$ for some p, forming the Bernstein-Dirichlet prior.

Bernstein densities have been used before.

1. **Maximum likelihood:** Studied by Guan (2016)
2. **Classical method:** Put $\lambda_\nu = F\left(\frac{\nu+1}{m+1}\right) - F\left(\frac{\nu}{m+1}\right)$ for a distribution function F. Motivated by Bernstein's proof of the Weierstraß approximation theorem. (Vitale, 1975)
3. **Non-parametric Bayes:** Petrone (1999) allowed $F \sim \mathrm{DP}(F_0, \alpha)$ and $m \sim p$ for some p, forming the Bernstein-Dirichlet prior.
4. **More!** E.g. shape constrained regression and quantile functions.

Definition of the polygram

Definition

A Bernstein polygram is a density $h(x \mid \mathbf{w}, \mathbf{s})$ of the form

$$h\left(x \mid \mathbf{w}, \mathbf{s}\right) = \sum_{k=1}^{|\mathbf{s}|-1} \sum_{\nu=0}^{m_k} w_{\nu k} b_{\nu, m_k} \left(\frac{x - s_{k-1}}{s_k - s_{k-1}} \right) 1_{[s_{k-1}, s_k)} \left(x\right)$$

Definition of the polygram

Definition

A Bernstein polygram is a density $h(x \mid \mathbf{w}, \mathbf{s})$ of the form

$$h\left(x \mid \mathbf{w}, \mathbf{s}\right) = \sum_{k=1}^{|\mathbf{s}|-1} \sum_{\nu=0}^{m_k} w_{\nu k} b_{\nu, m_k} \left(\frac{x - s_{k-1}}{s_k - s_{k-1}} \right) 1_{[s_{k-1}, s_k)} \left(x\right)$$

- $\mathbf{w}$: A list of weight vectors. It sums to 1.

Definition of the polygram

Definition

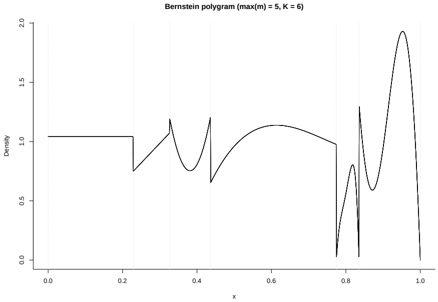
A Bernstein polygram is a density $h(x \mid \mathbf{w}, \mathbf{s})$ of the form

$$h\left(x \mid \mathbf{w}, \mathbf{s}\right) = \sum_{k=1}^{|\mathbf{s}|-1} \sum_{\nu=0}^{m_k} w_{\nu k} b_{\nu, m_k} \left(\frac{x - s_{k-1}}{s_k - s_{k-1}} \right) 1_{[s_{k-1}, s_k)} \left(x\right)$$

- $\mathbf{w}$: A list of weight vectors. It sums to 1.
- $\mathbf{s}$: A vector of split points $s_0 < s_1 < \cdots < s_{|\mathbf{s}|-1}$. So $[s_0, s_{|\mathbf{s}|-1}]$ is the support of h.

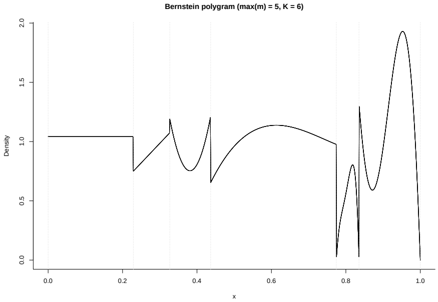
This is what a polygram can look like!

- Breathtaking density with unevenen **s** and $m_k = k - 1$.



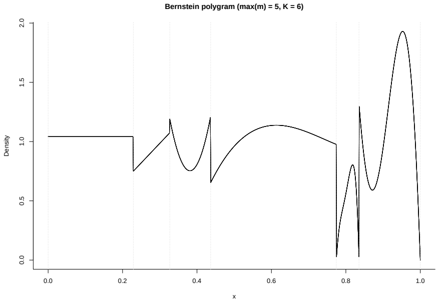
This is what a polygram can look like!

- Breathtaking density with unevenen **s** and $m_k = k - 1$.
- Looks pretty useless — discontinuous and ugly.



This is what a polygram can look like!

- Breathtaking density with unevenen **s** and $m_k = k - 1$.
- Looks pretty useless — discontinuous and ugly.
- Time to connect them as they do with splines!



How to make a spline from a polygram

- Let p_k a natural number so that the Bernstein polygram has continuous derivatives up to order p_k at the kth split.

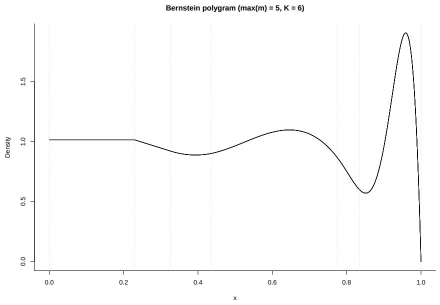
How to make a spline from a polygram

- Let p_k a natural number so that the Bernstein polygram has continuous derivatives up to order p_k at the kth split.
- $p_k \geq -1$ for all k, always

How to make a spline from a polygram

- Let p_k a natural number so that the Bernstein polygram has continuous derivatives up to order p_k at the kth split.
- $p_k \geq -1$ for all k, always
- Enforcing $p_k \geq 0$ involves only linear constraints in $\mathbf{w}$, with easily computable coefficients!

How to make a spline from a polygram



- Let p_k a natural number so that the Bernstein polygram has continuous derivatives up to order p_k at the kth split.
- $p_k \geq -1$ for all k, always
- Enforcing $p_k \geq 0$ involves only linear constraints in $\mathbf{w}$, with easily computable coefficients!

Don’t be unfair, minimize the square

- Maximum likelihood is unnatural to use in this problem; computationally hard.

Don't be unfair, minimize the square

- Maximum likelihood is unnatural to use in this problem; computationally hard.
- But least squares, or L_2-minimisation is a quadratic program with linear constraints!

$$\begin{array}{ll} \min & c\mathbf{w} + \frac{1}{2}\mathbf{w}^T Q\mathbf{w} \\ \text{s.t.} & A\mathbf{w} \leq b \end{array}$$

Don't be unfair, minimize the square

- Maximum likelihood is unnatural to use in this problem; computationally hard.
- But least squares, or L_2-minimisation is a quadratic program with linear constraints!
- Many efficient premade solvers: `quadprog`, `mosek`, etc.

$$\begin{array}{ll} \min & c\mathbf{w} + \frac{1}{2}\mathbf{w}^T Q\mathbf{w} \\ \text{s.t.} & A\mathbf{w} \leq b \end{array}$$

Don't be unfair, minimize the square

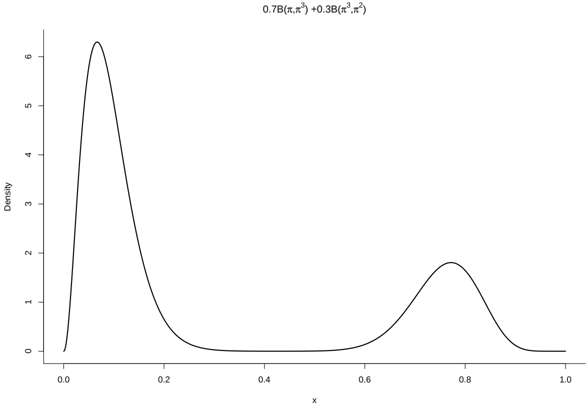
- Maximum likelihood is unnatural to use in this problem; computationally hard.
- But least squares, or L_2-minimisation is a quadratic program with linear constraints!
- Many efficient premade solvers: `quadprog`, `mosek`, etc.

$$\begin{array}{ll} \min & c\mathbf{w} + \frac{1}{2}\mathbf{w}^T Q\mathbf{w} \\ \text{s.t.} & A\mathbf{w} \leq b \end{array}$$

- c involves only "local" moments — easy to use something else than the ecdf. Q even simpler.

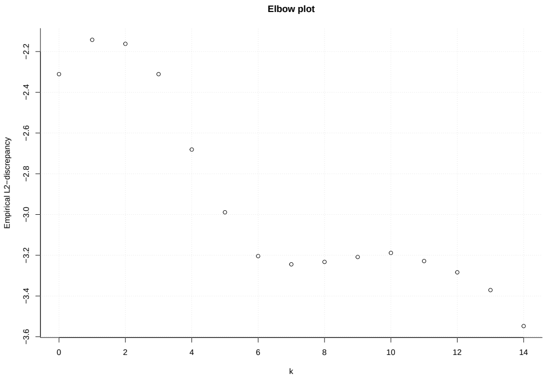
A bimodal mixture of Betas

What is $0.7\mathrm{Beta}(\pi, \pi^3) + 0.3\mathrm{Beta}(\pi^3, \pi^2)$?



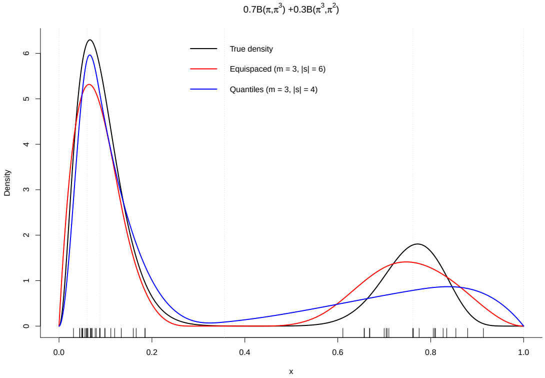
How to select **s** and **m**? The elbow plot!

Here $m = 3$ and $p = 2$, equispaced bins.



Estimated density together with the real density.

Looks good! Equispaced does better than quantiles.



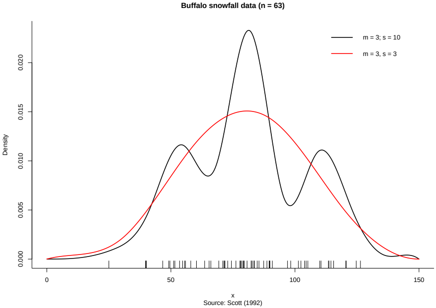
Is the density trimodal or unimodal?

- Data from Scott's *Multivariate Density Estimation*, 1992.

Is the density trimodal or unimodal?

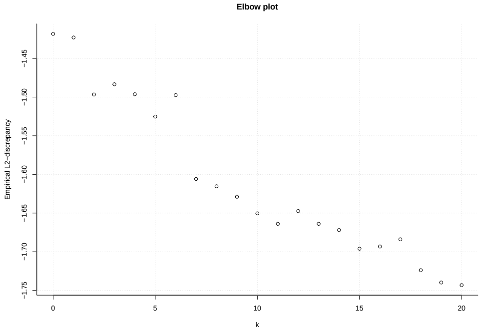
- Data from Scott's *Multivariate Density Estimation*, 1992.
- Annual snowfall in Buffalo, NY from 1910 to 1972.

Is the density trimodal or unimodal?

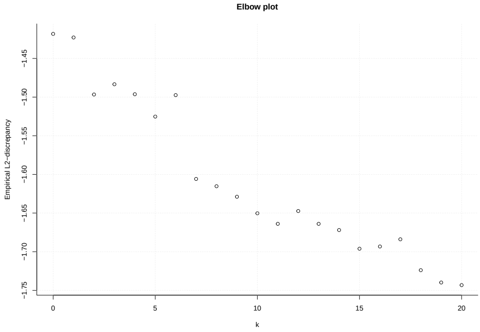


- Data from Scott's *Multivariate Density Estimation*, 1992.
- Annual snowfall in Buffalo, NY from 1910 to 1972.
- Unimodal or trimodal?

An elbow plot of the Buffalo snowfall data



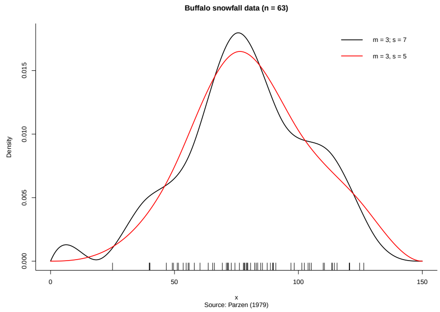
An elbow plot of the Buffalo snowfall data



Supports both 5 and 7 split points!

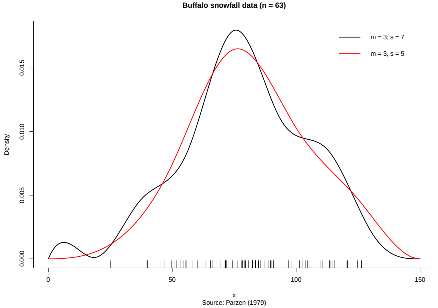
The post-elbow plot plot of the Buffalo snowfall data

- The density looks unimodal!



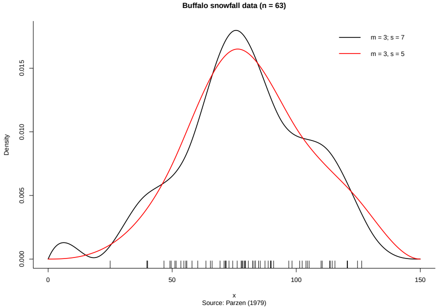
The post-elbow plot plot of the Buffalo snowfall data

- The density looks unimodal!
- Note the bump in the left tail.



The post-elbow plot plot of the Buffalo snowfall data

- The density looks unimodal!
- Note the bump in the left tail.
- Can be "fixed" by reducing connectedness constraints.



Berstein polygrams admit a large variety of shape constraints

- Monotone densities, Grenander (1956)

Berstein polygrams admit a large variety of shape constraints

- Monotone densities, Grenander (1956)
- Convex decreasing densities, Anevski (1994)

Berstein polygrams admit a large variety of shape constraints

- Monotone densities, Grenander (1956)
- Convex decreasing densities, Anevski (1994)
- Log-concave densities, Dümbgen and Rufibach (2009) - logsplines!

Berstein polygrams admit a large variety of shape constraints

- Monotone densities, Grenander (1956)
- Convex decreasing densities, Anevski (1994)
- Log-concave densities, Dümbgen and Rufibach (2009) - logsplines!
- Resource: Groeneboom and Jongbloed's *Nonparametric Estimation under Shape Constraints*, 2014.

Berstein polygrams admit a large variety of shape constraints

- Monotone densities, Grenander (1956)
- Convex decreasing densities, Anevski (1994)
- Log-concave densities, Dümbgen and Rufibach (2009) - logsplines!
- Resource: Groeneboom and Jongbloed's *Nonparametric Estimation under Shape Constraints*, 2014.
- Easy to implement shape constraints for Bernstein polygrams:

Berstein polygrams admit a large variety of shape constraints

- Monotone densities, Grenander (1956)
- Convex decreasing densities, Anevski (1994)
- Log-concave densities, Dümbgen and Rufibach (2009) - logsplines!
- Resource: Groeneboom and Jongbloed's *Nonparametric Estimation under Shape Constraints*, 2014.
- Easy to implement shape constraints for Bernstein polygrams:
    - Convex / concave, monotone, symmetric

Berstein polygrams admit a large variety of shape constraints

- Monotone densities, Grenander (1956)
- Convex decreasing densities, Anevski (1994)
- Log-concave densities, Dümbgen and Rufibach (2009) - logsplines!
- Resource: Groeneboom and Jongbloed's *Nonparametric Estimation under Shape Constraints*, 2014.
- Easy to implement shape constraints for Bernstein polygrams:
  - Convex / concave, monotone, symmetric
  - Specify quantiles and moments

Berstein polygrams admit a large variety of shape constraints

- Monotone densities, Grenander (1956)
- Convex decreasing densities, Anevski (1994)
- Log-concave densities, Dümbgen and Rufibach (2009) - logsplines!
- Resource: Groeneboom and Jongbloed's *Nonparametric Estimation under Shape Constraints*, 2014.
- Easy to implement shape constraints for Bernstein polygrams:
    - Convex / concave, monotone, symmetric
    - Specify quantiles and moments
    - Unimodal, bimodal, trimodal, etc. (iterative procedure)

Berstein polygrams generalize Grenander's estimator

- Least squares solution to a decreasing density on $[0, c)$.

Berstein polygrams generalize Grenander's estimator

- Least squares solution to a decreasing density on $[0, c)$.
  - The slope of the least concave majorant of the empirical density function.

Berstein polygrams generalize Grenander's estimator

- Least squares solution to a decreasing density on $[0, c)$.
  - The slope of the least concave majorant of the empirical density function.
  - A special case of Bernstein polygrams with $m = 0$ and decreasing constraints in effect.

Berstein polygrams generalize Grenander's estimator

- Least squares solution to a decreasing density on $[0, c)$.
  - The slope of the least concave majorant of the empirical density function.
  - A special case of Bernstein polygrams with $m = 0$ and decreasing constraints in effect.
- As $|s| \to \infty$, any decreasing Bernstein polygram (with fixed m) will converge pointwise to the Grenander estimator.

Berstein polygrams generalize Grenander's estimator

- Least squares solution to a decreasing density on $[0, c)$.
  - The slope of the least concave majorant of the empirical density function.
  - A special case of Bernstein polygrams with $m = 0$ and decreasing constraints in effect.
- As $|s| \to \infty$, any decreasing Bernstein polygram (with fixed m) will converge pointwise to the Grenander estimator.
  - Question: Are there natural addition constraints, e.g. on the derivatives, to ensure least squares solutions for $m > 0$?

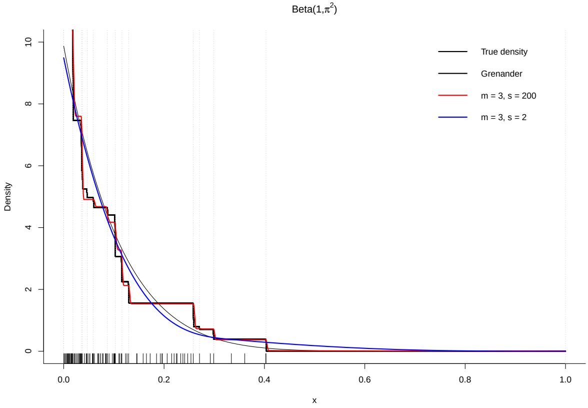
Berstein polygrams generalize Grenander's estimator

- Least squares solution to a decreasing density on $[0, c)$.
  - The slope of the least concave majorant of the empirical density function.
  - A special case of Bernstein polygrams with $m = 0$ and decreasing constraints in effect.
- As $|s| \to \infty$, any decreasing Bernstein polygram (with fixed m) will converge pointwise to the Grenander estimator.
  - Question: Are there natural addition constraints, e.g. on the derivatives, to ensure least squares solutions for $m > 0$?
  - Appears to work well as an estimator of monotone densities - but needs a method for selecting s!

Berstein polygrams generalize Grenander's estimator

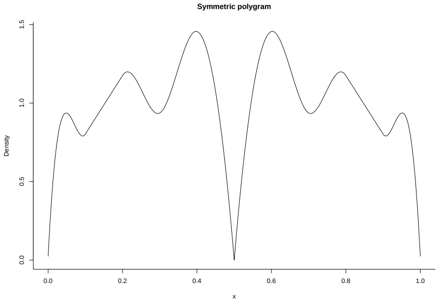
- Least squares solution to a decreasing density on $[0, c)$.
  - The slope of the least concave majorant of the empirical density function.
  - A special case of Bernstein polygrams with $m = 0$ and decreasing constraints in effect.
- As $|s| \to \infty$, any decreasing Bernstein polygram (with fixed m) will converge pointwise to the Grenander estimator.
  - Question: Are there natural addition constraints, e.g. on the derivatives, to ensure least squares solutions for $m > 0$?
  - Appears to work well as an estimator of monotone densities - but needs a method for selecting s!
- Same results and questions for decreasing convex densities.

Example of decreasing density



Good densities are hard to find

```
set.seed(1984)
x = runif(11)
m = c(3,1,3,3,2,2,3,3,1,3)
p =  c(1,1,2,1,0,1,2,1,1)
obj = polygram(x ~ symmetric,
 m = m, p = p)
plot(obj, bins = FALSE,
 main = "Symmetric polygram")
```



Download the R-package!

```
install.packages(devtools)
library(devtools)
install_github("JonasMoss/polygrams")

# Use example with 6 split points and m = 3, p = 2:
x = rbeta(200, 2, 7)
polygram_object = polygram(x, s = 6))
plot(polygram_object)

# Example for a convex and decreasing density
x = rbeta(200, 1, 9)
polygram_object = polygram(x ~ convex + decreasing, s = 5)
```