

Testing for Log-Concavity of Densities

Martin Hazelton

`m.hazelton@massey.ac.nz`

Massey University

29 August 2011

Concave Functions

Definition

A function $f: \mathcal{X} \rightarrow \mathbb{R}$ is **concave** iff for any $\theta \in [0, 1]$, $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X} \subseteq \mathbb{R}^d$,

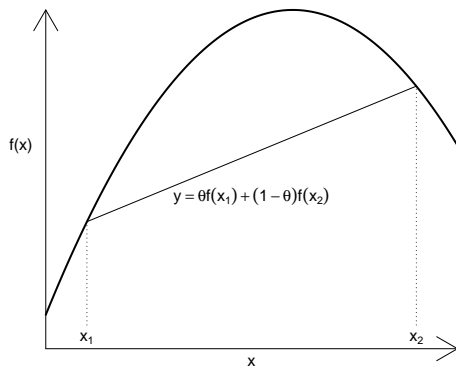
$$f(\theta \mathbf{x}_1 + (1 - \theta) \mathbf{x}_2) \geq \theta f(\mathbf{x}_1) + (1 - \theta) f(\mathbf{x}_2).$$

Concave Functions

Definition

A function $f: \mathcal{X} \rightarrow \mathbb{R}$ is **concave** iff for any $\theta \in [0, 1]$, $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X} \subseteq \mathbb{R}^d$,

$$f(\theta \mathbf{x}_1 + (1 - \theta) \mathbf{x}_2) \geq \theta f(\mathbf{x}_1) + (1 - \theta) f(\mathbf{x}_2).$$



If f is twice differentiable:

- f concave $\Leftrightarrow f'' \leq 0$ ($d = 1$);
- f concave $\Leftrightarrow Hf \preceq 0$ ($d \geq 1$).

Log-Concave Densities

Definition

A density f is log-concave iff $\log(f)$ is concave.

Log-Concave Densities

Definition

A density f is log-concave iff $\log(f)$ is concave.

Example (Univariate densities)

- ✓ Normal; Gamma with shape parameter ≥ 1 ; Beta(α, β) with $\alpha, \beta > 1$.
- ✗ Cauchy; Pareto; log-normal.

Log-Concave Densities

Definition

A density f is log-concave iff $\log(f)$ is concave.

Example (Univariate densities)

- ✓ Normal; Gamma with shape parameter ≥ 1 ; Beta(α, β) with $\alpha, \beta > 1$.
- ✗ Cauchy; Pareto; log-normal.

Example (Multivariate mixture)

Multivariate normal mixture $\frac{1}{2}\phi(\mathbf{x}) + \frac{1}{2}\phi(\mathbf{x} - \boldsymbol{\mu})$ log-concave iff $\|\boldsymbol{\mu}\| \leq 2$.

We Care Because...?

Applications

- Economists like (log)-concave functions for modelling.
- Class of mixtures of log-concave densities is a flexible model.

We Care Because...?

Applications

- Economists like (log)-concave functions for modelling.
- Class of mixtures of log-concave densities is a flexible model.
- Not entirely convincing?

We Care Because...?

Applications

- Economists like (log)-concave functions for modelling.
- Class of mixtures of log-concave densities is a flexible model.
- Not entirely convincing?

Theory

- Log-concave densities can be estimated by nonparametric maximum likelihood estimation.
- Shape constraint prevents degeneracy.

We Care Because...?

Applications

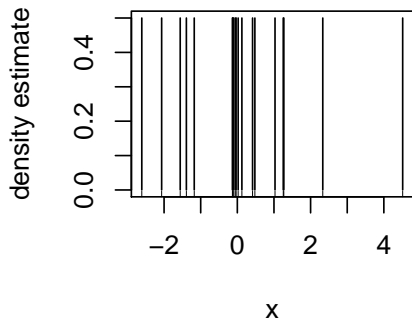
- Economists like (log)-concave functions for modelling.
- Class of mixtures of log-concave densities is a flexible model.
- Not entirely convincing?

Theory

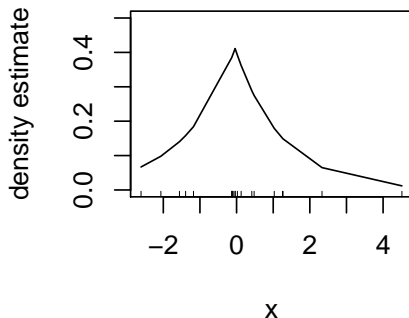
- Log-concave densities can be estimated by nonparametric maximum likelihood estimation.
- Shape constraint prevents degeneracy.
- Interesting!

Nonparametric ML Estimation of a Density

No shape constraint



Log-concave



Testing for Log-Concavity

- $\mathbf{x}_1, \dots, \mathbf{x}_n$ a d -dimensional random sample from f .
- Test H_0 : f is log-concave on $\tilde{\mathcal{X}} \subseteq \mathcal{X}$.
 - ▶ Not enough information in tails to assess log-concavity there.
 - ▶ Focus on central region $\tilde{\mathcal{X}}$ containing most (e.g. 90%) of data.

Testing for Log-Concavity

- $\mathbf{x}_1, \dots, \mathbf{x}_n$ a d -dimensional random sample from f .
- Test H_0 : f is log-concave on $\tilde{\mathcal{X}} \subseteq \mathcal{X}$.
 - ▶ Not enough information in tails to assess log-concavity there.
 - ▶ Focus on central region $\tilde{\mathcal{X}}$ containing most (e.g. 90%) of data.
- Several tests in literature for case $d = 1$.
- Multivariate setting has proven more challenging.

Kernel Density Estimation and Log-Concavity

Definition

The **kernel density estimate** constructed from $\mathbf{x}_1, \dots, \mathbf{x}_n$ is:

$$\hat{f}(\mathbf{x}|h) = n^{-1} \sum_{i=1}^n K_h(\mathbf{x} - \mathbf{x}_i)$$

- $K_h(\mathbf{x}) = h^{-1}K(\mathbf{x}/h)$, where K is isotropic kernel: we take $K = \phi$;
- h is the bandwidth, controlling degree of smoothing.

Kernel Density Estimation and Log-Concavity

Definition

The **kernel density estimate** constructed from $\mathbf{x}_1, \dots, \mathbf{x}_n$ is:

$$\hat{f}(\mathbf{x}|h) = n^{-1} \sum_{i=1}^n K_h(\mathbf{x} - \mathbf{x}_i)$$

- $K_h(\mathbf{x}) = h^{-1}K(\mathbf{x}/h)$, where K is isotropic kernel: we take $K = \phi$;
- h is the bandwidth, controlling degree of smoothing.

Theorem

- 1 $\hat{f}(\cdot|h)$ is log-concave for sufficiently large h .
- 2 If $\hat{f}(\cdot|h_0)$ is log-concave, then so is $\hat{f}(\cdot|h)$ for all $h \geq h_0$.

A Test Statistic

What the Theorem says:

- $\hat{f}(\cdot|h)$ becomes more log-concave as h increases.
- There exists smallest bandwidth, h_0 , to make \hat{f} log-concave.

A Test Statistic

What the Theorem says:

- $\hat{f}(\cdot|h)$ becomes more log-concave as h increases.
- There exists smallest bandwidth, h_0 , to make \hat{f} log-concave.

Test design:

- Use h_0 as **test statistic**.
- Will reject H_0 if h_0 too large.

Smooth Bootstrap P-values

Compute p-value for observed test statistic h_0 by simulation.

Smooth Bootstrap P-values

Compute p-value for observed test statistic h_0 by simulation.

Algorithm

- 1 Draw random (smooth bootstrap) sample $\mathbf{x}_1^\dagger, \dots, \mathbf{x}_n^\dagger$ from $\hat{f}(\cdot | h_0)$.
- 2 Rescale to $\mathbf{x}_1^*, \dots, \mathbf{x}_n^*$ so that mean vector and covariance matrix match observed data.
 - ▶ Counters variance inflation.
- 3 Steps 1 and 2 repeated N times. From i bootstrap sample let h_i^* be critical bandwidth
- 4 Compute bootstrap p-value $p = N^{-1} \sum_{i=1}^N I_{\{h_i^* > h_0\}}$.

Avoiding a Computational Nightmare

- Computation of critical bandwidths time consuming.
- But only need to do so for observed h_0 .
- Then check log-concavity of $f^*(\cdot|h_0)$ computed from $\mathbf{x}_1^*, \dots, \mathbf{x}_n^*$.
- $f^*(\cdot|h_0)$ log-concave iff $h^* \leq h_0$.
 - ▶ Remember Theorem.



Example: Testing for Breast Cancer

The Data

Full dataset

- Measurements taken on digitized images of fine needle aspirates of breast masses.
- 30 measurements on 569 subjects.
- For each breast mass a diagnosis of benign (357 cases) or malignant (212 cases) available.

Example: Testing for Breast Cancer

The Data

Full dataset

- Measurements taken on digitized images of fine needle aspirates of breast masses.
- 30 measurements on 569 subjects.
- For each breast mass a diagnosis of benign (357 cases) or malignant (212 cases) available.

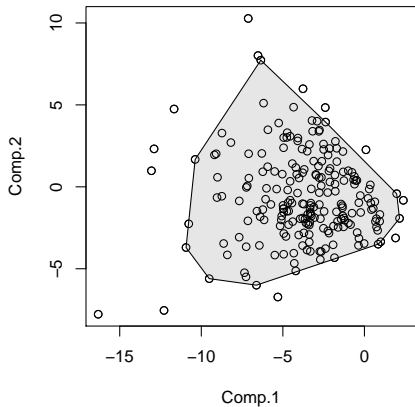
Dimension Reduction

- We will look at first two principal components, split by diagnosis.

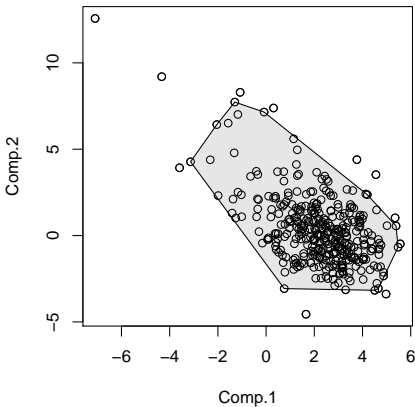
Example: Testing for Breast Cancer

Scatterplots of First Two Principal Components by Diagnosis

Malignant



Benign



Example: Testing for Breast Cancer

Testing for Log-Concavity

- Model proposed for these data using separate log-concave densities for benign and malignant samples.
- Test for log-concavity of these densities.

Malignant data

critical bandwidth $h_0 = 2.83$ with p-value 0.645.

Benign data

critical bandwidth $h_0 = 2.97$ with p-value 0.00.

Some References to Finish With

- Copy of these slides

<http://www.massey.ac.nz/~mhazelto/seminars>

- Cule, M., Gramacy, R., and Samworth, R. (2009). LogConcDEAD: An R package for maximum likelihood estimation of a multivariate log-concave density, *Journal of Statistical Software* **29** (2).
- Cule, M., Samworth, R., Stewart, M. (2010). Maximum likelihood estimation of a multi-dimensional log-concave distribution (with discussion). *Journal of the Royal Statistical Society Series B* **72**, 545-607.
- Hazelton, M.L. (2011). Assessing log-concavity of multivariate densities. *Statistics and Probability Letters* **81**, 121–125.