Asymptotic Theory for Density Ridges

Yen-Chi Chen

Christopher R. Genovese   Larry Wasserman

Department of Statistics
Carnegie Mellon University

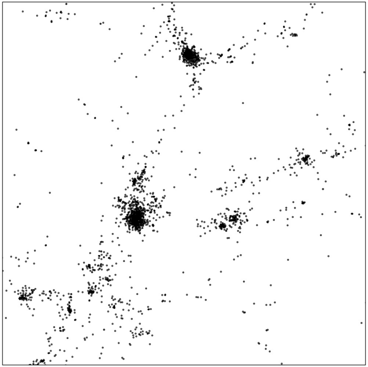
November 14, 2015

Density Ridges: High Density Curves

Density ridges are curves characterizing high density regions.

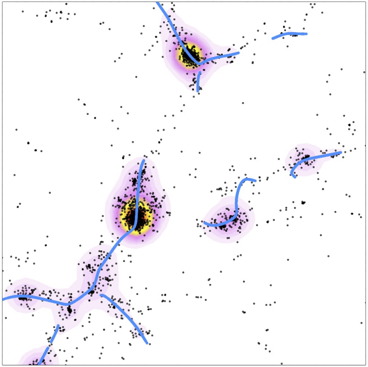
Density Ridges: High Density Curves

Density ridges are curves characterizing high density regions.

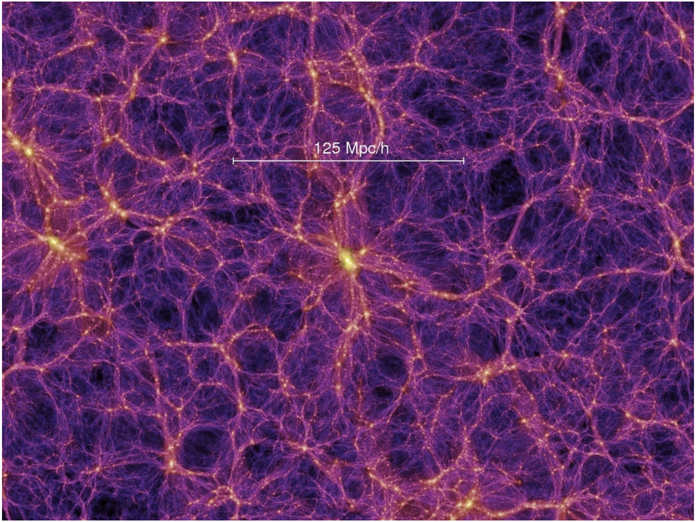


Density Ridges: High Density Curves

Density ridges are curves characterizing high density regions.

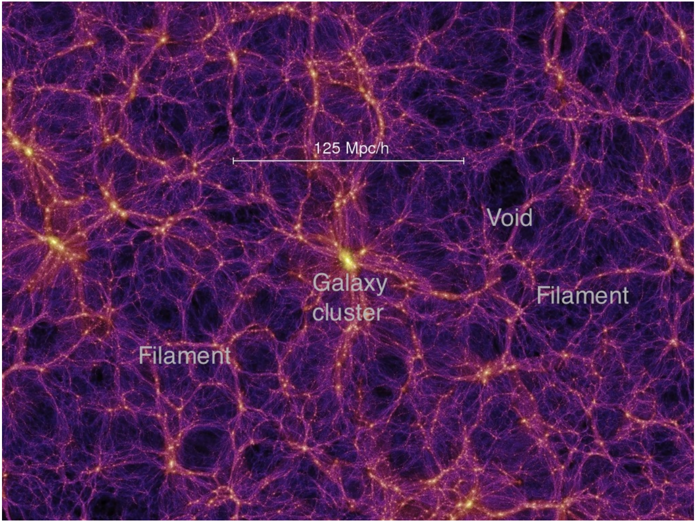


Application of Ridges: Cosmology



Credit: Millennium Simulation

Application of Ridges: Cosmology



Credit: Millennium Simulation

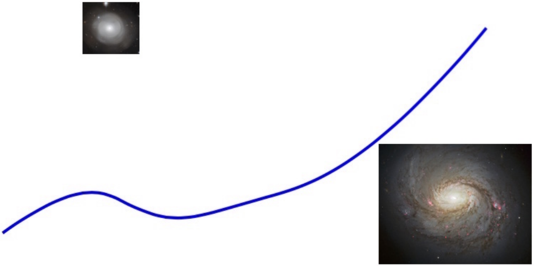
The Importance of Filaments

Cosmic filaments play key roles in astronomy research.

The Importance of Filaments

Cosmic filaments play key roles in astronomy research.

- A galaxy's color, mass, and size are associated with filaments.

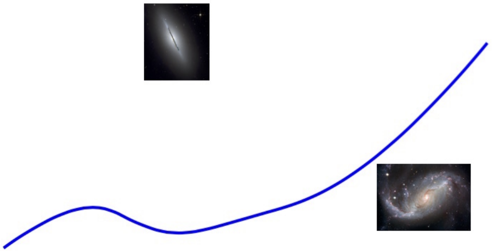


→ **Chen** et al. 'Detecting Effects of Filaments on Galaxy Properties in Sloan Digital Sky Survey III' (2015)

The Importance of Filaments

Cosmic filaments play key roles in astronomy research.

- A galaxy's color, mass, and size are associated with filaments.
- A galaxy's shape is associated with filaments.

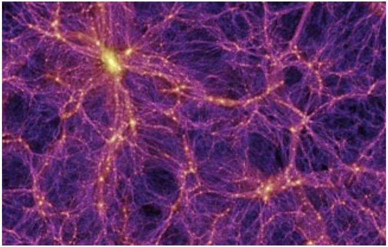


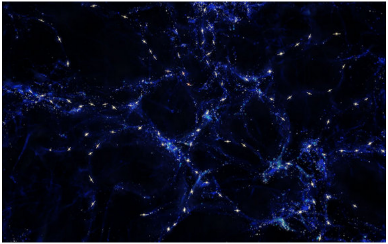
→ **Chen** et al. 'Investigating Galaxy-Filament Alignment in Hydrodynamic Simulations using Density Ridges' (Mon. Not. Roy. Astro. Soc. 2015)

The Importance of Filaments

Cosmic filaments play key roles in astronomy research.

- A galaxy's color, mass, and size are associated with filaments.
- A galaxy's shape is associated with filaments.
- Filaments can be used to constrain the cosmological models.





- Credit: Millennium Simulation and ESO/M. Kornmesser.

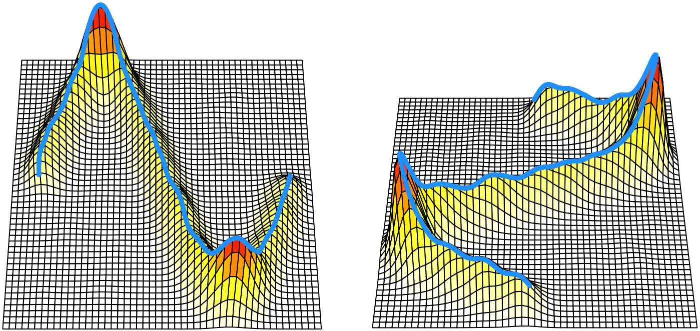
Density Ridges

A statistical model for filaments is the *density ridges*.

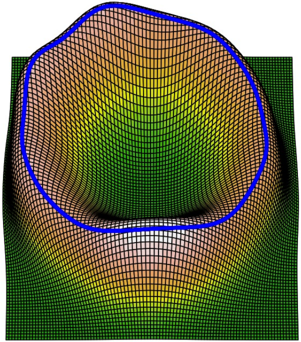
Example: Ridges in Mountains

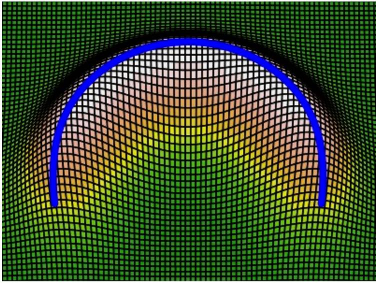
Credit: Google

Example: Ridges in Smooth Functions

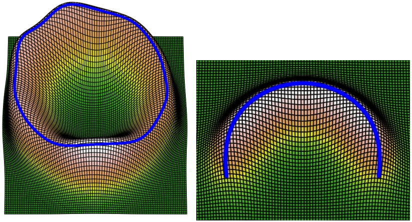


Example: Ridges in Smooth Functions



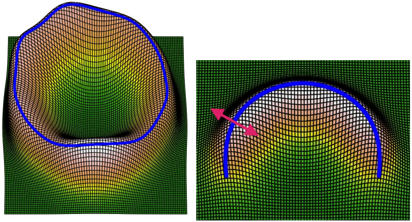


Ridges: Local Modes in Subspace



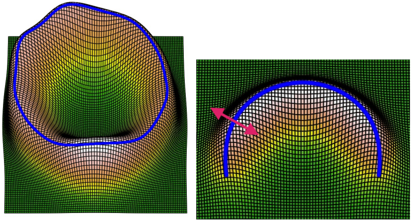
- A generalized local mode in a specific 'subspace'.

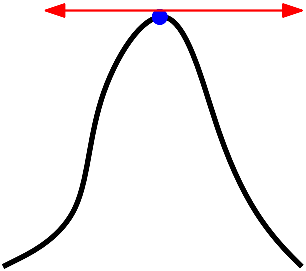
Ridges: Local Modes in Subspace



- A generalized local mode in a specific 'subspace'.

Ridges: Local Modes in Subspace





- A generalized local mode in a specific 'subspace'.

Formal Definition of Density Ridges

- $p(x)$: a density function.

Formal Definition of Density Ridges

- $p(x)$: a density function.
- $(\lambda_j(x), v_j(x))$: jth eigenvalue/vector of $H(x) = \nabla\nabla p(x)$.

Formal Definition of Density Ridges

- $p(x)$: a density function.
- $(\lambda_j(x), v_j(x))$: jth eigenvalue/vector of $H(x) = \nabla\nabla p(x)$.
- $V(x) = [v_2(x), \cdots, v_d(x)]$: matrix of 2nd to last eigenvectors.

Formal Definition of Density Ridges

- $p(x)$: a density function.
- $(\lambda_j(x), v_j(x))$: jth eigenvalue/vector of $H(x) = \nabla\nabla p(x)$.
- $V(x) = [v_2(x), \cdots, v_d(x)]$: matrix of 2nd to last eigenvectors.
- $V(x)V(x)^T$: a projection.

Formal Definition of Density Ridges

- $p(x)$: a density function.
- $(\lambda_j(x), v_j(x))$: jth eigenvalue/vector of $H(x) = \nabla\nabla p(x)$.
- $V(x) = [v_2(x), \cdots, v_d(x)]$: matrix of 2nd to last eigenvectors.
- $V(x)V(x)^T$: a projection.
- Ridges:

$$R = \text{Ridge}(p) = \{x : V(x)V(x)^T \nabla p(x) = 0, \lambda_2(x) < 0\},$$

Formal Definition of Density Ridges

- $p(x)$: a density function.
- $(\lambda_j(x), v_j(x))$: jth eigenvalue/vector of $H(x) = \nabla\nabla p(x)$.
- $V(x) = [v_2(x), \cdots, v_d(x)]$: matrix of 2nd to last eigenvectors.
- $V(x)V(x)^T$: a projection.
- Ridges:

$$R = \text{Ridge}(p) = \{x : V(x)V(x)^T \nabla p(x) = 0, \lambda_2(x) < 0\},$$

- Local modes:

$$\text{Mode}(p) = \{x : \nabla p(x) = 0, \lambda_1(x) < 0\}.$$

Estimator and Algorithm

We use the plug-in estimate:

$$\widehat{R}_n = \text{Ridge}(\widehat{p}_n),$$

where $\widehat{p}_n$ is the KDE.

Estimator and Algorithm

We use the plug-in estimate:

$$\widehat{R}_n = \text{Ridge}(\widehat{p}_n),$$

where $\widehat{p}_n$ is the KDE.

- In general, finding ridges from a given function is hard.

Estimator and Algorithm

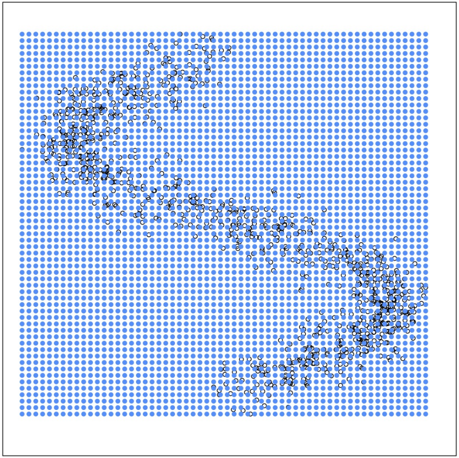
We use the plug-in estimate:

$$\widehat{R}_n = \text{Ridge}(\widehat{p}_n),$$

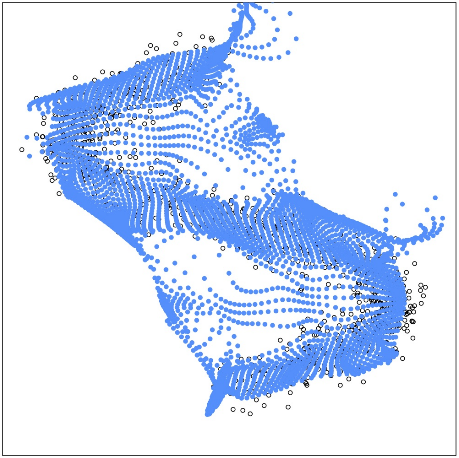
where $\widehat{p}_n$ is the KDE.

- In general, finding ridges from a given function is hard.
- The Subspace Constraint Mean Shift (SCMS; Ozertem2011) algorithm allows us to find $\widehat{R}_n$, ridges of the KDE.

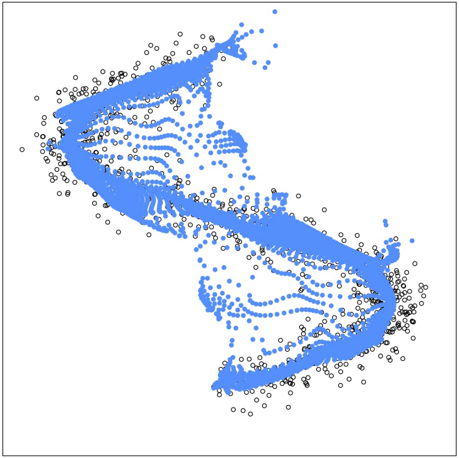
SCMS: Ridge Recovery Algorithm



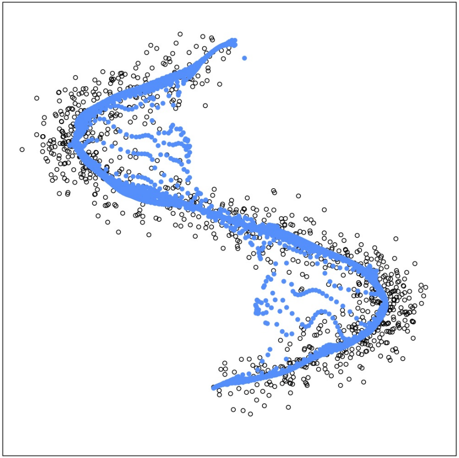
SCMS: Ridge Recovery Algorithm



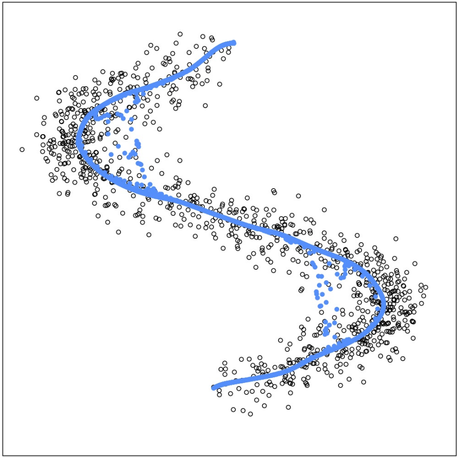
SCMS: Ridge Recovery Algorithm



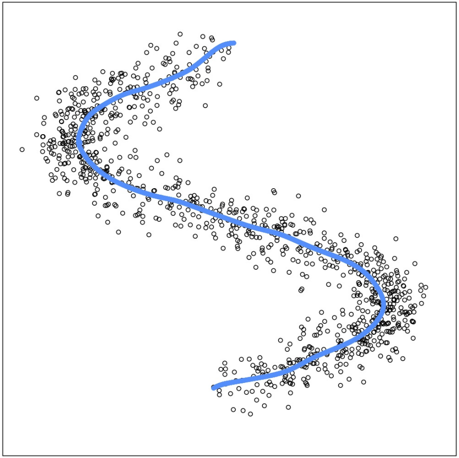
SCMS: Ridge Recovery Algorithm



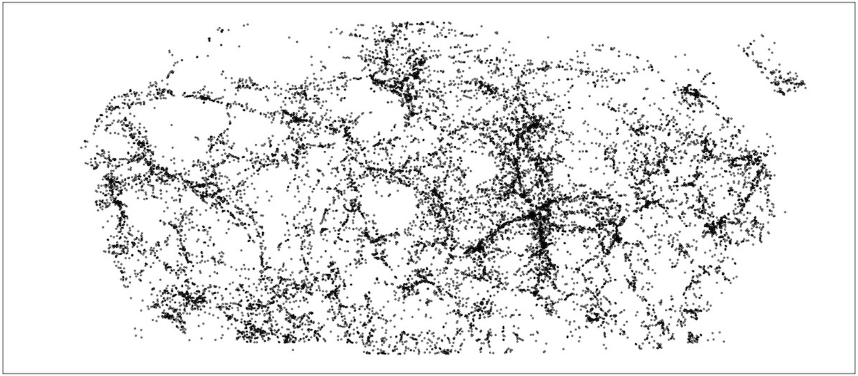
SCMS: Ridge Recovery Algorithm



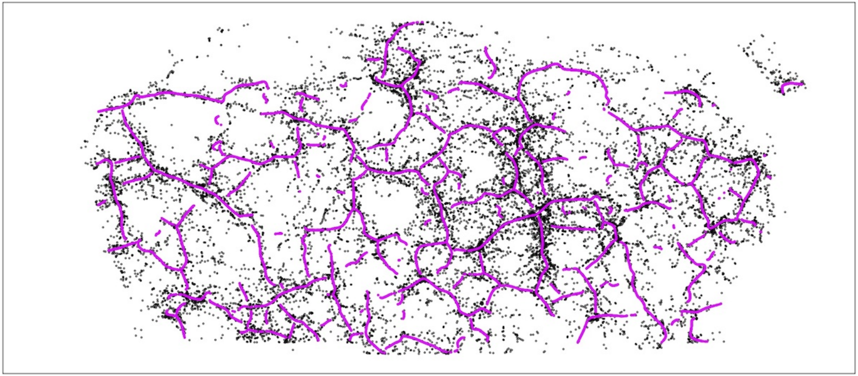
SCMS: Ridge Recovery Algorithm



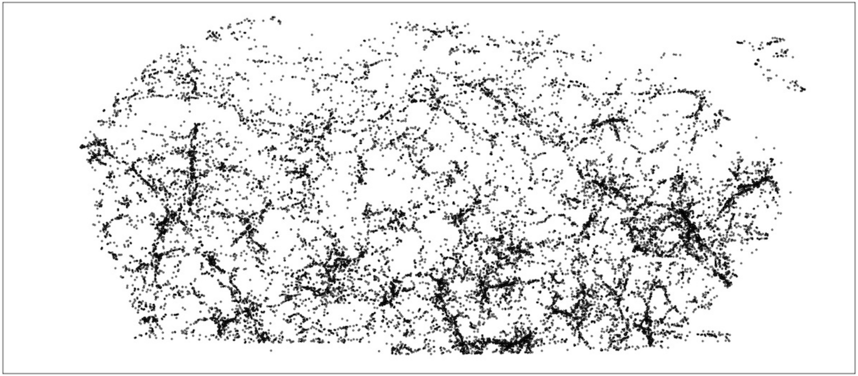
Example for Estimated Density Ridges



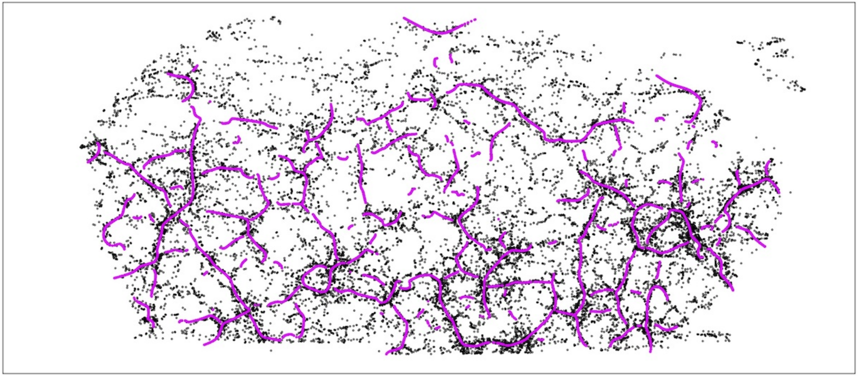
Example for Estimated Density Ridges



Example for Estimated Density Ridges



Example for Estimated Density Ridges



Statistical Inference: Confidence Sets

Having estimators is not enough for statistical inference.
We need confidence sets for density ridges.

Statistical Inference: Confidence Sets

Having estimators is not enough for statistical inference.
We need confidence sets for density ridges.
Namely, we want to find a set $C_{1-\alpha,n}$ from the data such that

$$\mathbb{P}\left(R \subset C_{1-\alpha,n}\right) \geq 1 - \alpha.$$

Statistical Inference: Confidence Sets

Having estimators is not enough for statistical inference.
We need confidence sets for density ridges.
Namely, we want to find a set $C_{1-\alpha,n}$ from the data such that

$$\mathbb{P}\left(R \subset C_{1-\alpha,n}\right) \geq 1-\alpha.$$

In what follows, we ignore the bias for estimating R and focus only on the stochastic variation of $\widehat{R}_n$.

Useful Metric: Hausdorff Distance

We introduce a useful metric–*the Hausdorff distance* for sets:

$$\mathrm{Haus}(A, B) = \max\left\{\sup_{x\in A} d(x, B), \sup_{x\in B} d(x, A)\right\},$$

where $d(x, A) = \inf_{y\in A} \|x - y\|$ is the projection distance.

Useful Metric: Hausdorff Distance

We introduce a useful metric–*the Hausdorff distance* for sets:

$$\mathrm{Haus}(A, B) = \max\left\{\sup_{x\in A} d(x, B), \sup_{x\in B} d(x, A)\right\},$$

where $d(x, A) = \inf_{y\in A} \|x - y\|$ is the projection distance.

- Haus is an $\mathcal{L}_\infty$ metric for sets.
- Consistency: $\mathrm{Haus}(\widehat{R}_n, R) = o_{\mathbb{P}}(1)$.

The $\oplus$ Operation

We define $A \oplus r = \{x : d(x, A) \leq r\}$.

The $\oplus$ Operation

We define $A \oplus r = \{x : d(x, A) \leq r\}$.

Then we have the following inclusion property:

$$A \subset B \oplus \mathrm{Haus}(A, B), \quad B \subset A \oplus \mathrm{Haus}(A, B).$$

Hausdorff Distance and Confidence Sets

We can use Hausdorff distance and $\oplus$ operation to construct confidence sets.
Let F_n be the CDF for $\text{Haus}(\widehat{R}_n, R)$ and $t_{1-\alpha} = F_n^{-1}(1-\alpha)$ be the $1-\alpha$ quantile.

Hausdorff Distance and Confidence Sets

We can use Hausdorff distance and $\oplus$ operation to construct confidence sets.
Let F_n be the CDF for $\mathrm{Haus}(\widehat{R}_n, R)$ and $t_{1-\alpha} = F_n^{-1}(1-\alpha)$ be the $1-\alpha$ quantile.

- It can be shown that

$$\mathbb{P}\left(R \subset \widehat{R}_n \oplus t_{1-\alpha}\right) \geq 1-\alpha.$$

$\rightarrow$ This follows from the property

$$A \subset B \oplus \mathrm{Haus}(A,B), \quad B \subset A \oplus \mathrm{Haus}(A,B).$$

Hausdorff Distance and Confidence Sets

We can use Hausdorff distance and $\oplus$ operation to construct confidence sets.
Let F_n be the CDF for $\mathsf{Haus}(\widehat{R}_n, R)$ and $t_{1-\alpha} = F_n^{-1}(1-\alpha)$ be the $1-\alpha$ quantile.

- It can be shown that

$$\mathbb{P}\left(R \subset \widehat{R}_n \oplus t_{1-\alpha}\right) \geq 1-\alpha.$$

$\rightarrow$ This follows from the property

$$A \subset B \oplus \mathsf{Haus}(A,B), \quad B \subset A \oplus \mathsf{Haus}(A,B).$$

- We need to find the distribution F_n.

Asymptotic Theory

Key observation:

$$\sqrt{nh^{d+2}}\text{Haus}(\widehat{R}_n, R) \approx \sqrt{nh^{d+2}} \sup_{x \in R} d(x, \widehat{R}_n)$$
$$\approx \sup \{\text{Empirical process on } R\}$$
$$\approx \sup \{\text{Gaussian process on } R\}.$$

Asymptotic Theory

Key observation:

$$\begin{aligned}\sqrt{nh^{d+2}}\text{Haus}(\widehat{R}_n, R) &\approx \sqrt{nh^{d+2}} \sup_{x\in R} d(x, \widehat{R}_n)\\ &\approx \sup\{\text{Empirical process on } R\}\\ &\approx \sup\{\text{Gaussian process on } R\}.\end{aligned}$$

Theorem

Under regularity conditions, there exists a tight Gaussian process $\mathbb{B}$ defined on a certain function space $\mathcal{F}$ such that

$$\sup_t \left| \mathbb{P}\left(\sqrt{nh^{d+2}}\text{Haus}(\widehat{R}_n, R) < t\right) - \mathbb{P}\left(\sup_{f\in\mathcal{F}} |\mathbb{B}(f)| < t\right)\right| = O\left(\left(\frac{\log^7 n}{nh^{d+2}}\right)^{1/8}\right).$$

The Bootstrap

Good news: we have the asymptotic behavior.
Bad news: the asymptotic behavior is complicated.

The Bootstrap

Good news: we have the asymptotic behavior.

Bad news: the asymptotic behavior is complicated.

$\longrightarrow$ A solution: the bootstrap.

The Bootstrap Consistency

- Bootstrap sample $\Longrightarrow$ bootstrap ridges $\widehat{R}_n^*$.
- Compute Haus$(\widehat{R}_n^*, \widehat{R}_n)$ to get a CDF estimator $\widehat{F}_n$.
- Choose $\widehat{t}_{1-\alpha}$ be the $1-\alpha$ quantile for $\widehat{F}_n$.

The Bootstrap Consistency

- Bootstrap sample $\Longrightarrow$ bootstrap ridges $\widehat{R}_n^*$.
- Compute Haus$(\widehat{R}_n^*, \widehat{R}_n)$ to get a CDF estimator $\widehat{F}_n$.
- Choose $\widehat{t}_{1-\alpha}$ be the $1-\alpha$ quantile for $\widehat{F}_n$.

It can be shown that

$$\begin{aligned}\sqrt{nh^{d+2}}\text{Haus}(\widehat{R}_n^*, \widehat{R}_n) &\approx \sup\left\{\text{Gaussian process on } \widehat{R}_n\right\}\\ &\approx \sup\left\{\text{Gaussian process on } R\right\}\\ &\approx \sqrt{nh^{d+2}}\text{Haus}(\widehat{R}_n, R).\end{aligned}$$

The Bootstrap Consistency

- Bootstrap sample $\Longrightarrow$ bootstrap ridges $\widehat{R}_n^*$.
- Compute $\text{Haus}(\widehat{R}_n^*, \widehat{R}_n)$ to get a CDF estimator $\widehat{F}_n$.
- Choose $\widehat{t}_{1-\alpha}$ be the $1-\alpha$ quantile for $\widehat{F}_n$.

It can be shown that

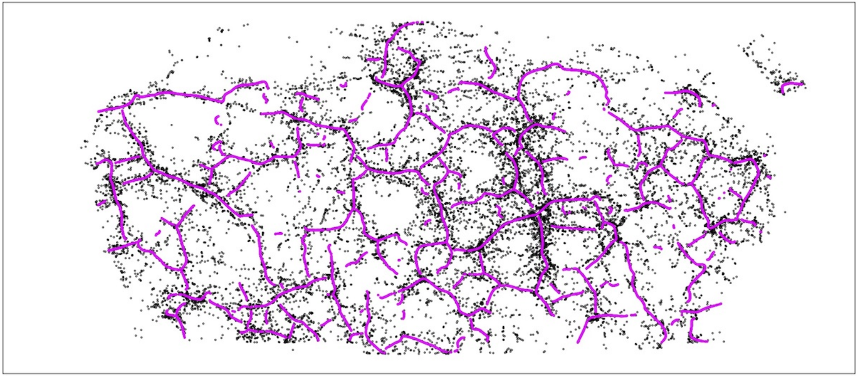
$$\begin{aligned}\sqrt{nh^{d+2}}\text{Haus}(\widehat{R}_n^*, \widehat{R}_n) &\approx \sup\left\{\text{Gaussian process on } \widehat{R}_n\right\}\\ &\approx \sup\left\{\text{Gaussian process on } R\right\}\\ &\approx \sqrt{nh^{d+2}}\text{Haus}(\widehat{R}_n, R).\end{aligned}$$

Theorem

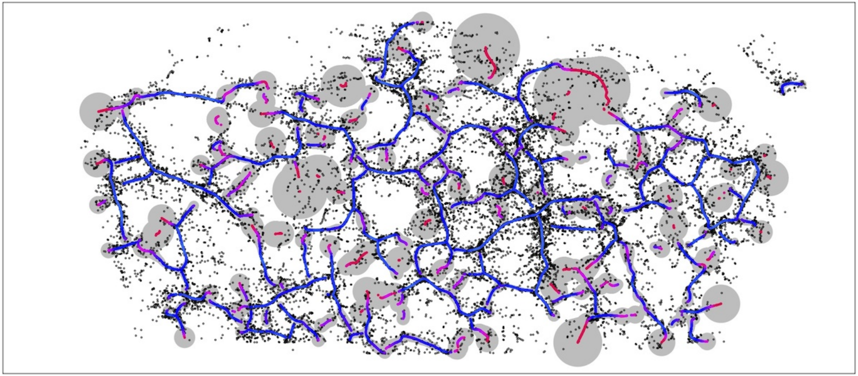
Under regularity conditions,

$$\mathbb{P}\left(R \subset \widehat{R}_n \oplus \widehat{t}_{1-\alpha}\right) = 1 - \alpha + O\left(\left(\frac{\log^7 n}{nh^{d+2}}\right)^{1/8}\right).$$

Example for Confidence Sets



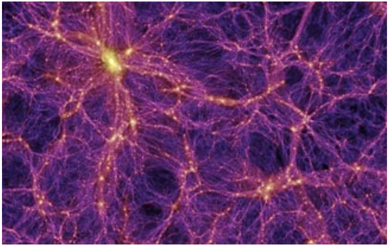
Example for Confidence Sets



Concluding Remarks

Density ridges are very cool objects because

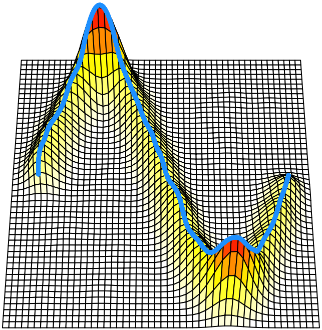
1. they have cosmological applications,



Concluding Remarks

Density ridges are very cool objects because

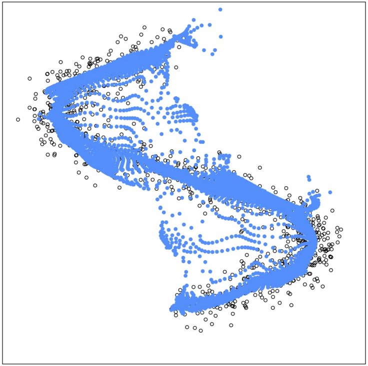
1. they have cosmological applications,
2. they are well-defined objects,



Concluding Remarks

Density ridges are very cool objects because

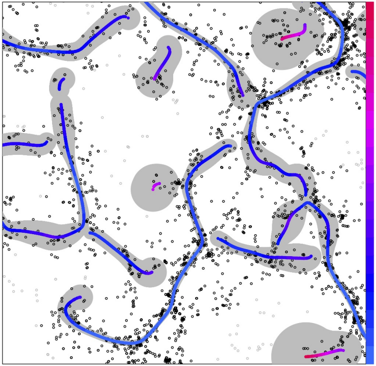
1. they have cosmological applications,
2. they are well-defined objects,
3. there is a fast algorithm to compute them,



Concluding Remarks

Density ridges are very cool objects because

1. they have cosmological applications,
2. they are well-defined objects,
3. there is a fast algorithm to compute them,
4. their statistical properties are well-studied.



Thank you!

References

1. Chen, Yen-Chi, Christopher R. Genovese, and Larry Wasserman. "Density Level Sets: Asymptotics, Inference, and Visualization." Submitted to the Journal of American Statistical Association. arXiv preprint arXiv:1504.05438 (2015).
2. Chen, Yen-Chi, Christopher R. Genovese, and Larry Wasserman. "Asymptotic theory for density ridges." To appear in the Annals of Statistics. arXiv preprint arXiv:1406.5663 (2014).
3. Chen, Yen-Chi, Christopher R. Genovese, Ryan J. Tibshirani, and Larry Wasserman. "Nonparametric Modal Regression." Under review of the Annals of Statistics. arXiv preprint arXiv:1412.1716 (2014).
4. Chernozhukov, Victor, Denis Chetverikov, and Kengo Kato. "Gaussian approximation of suprema of empirical processes." The Annals of Statistics 42, no. 4 (2014): 1564-1597.
5. Chernozhukov, Victor, Denis Chetverikov, and Kengo Kato. "Anti-concentration and honest, adaptive confidence bands." The Annals of Statistics 42, no. 5 (2014): 1787-1818.
6. Einbeck, Jochen, and Gerhard Tutz. "Modelling beyond regression functions: an application of multimodal regression to speedflow data." Journal of the Royal Statistical Society: Series C (Applied Statistics) 55, no. 4 (2006): 461-475.
7. Genovese, Christopher R., et al. "Nonparametric ridge estimation." The Annals of Statistics 42.4 (2014): 1511-1545.
8. Ozertem, Umut, and Deniz Erdogmus. "Locally defined principal curves and surfaces." The Journal of Machine Learning Research 12 (2011): 1249-1286.

Smoothed Density Ridges

In particular, we focus on making inference for the smoothed version of the density, denoted as p_h:

$$p_h(x) = p \otimes K_h(x) = \mathbb{E}\left(\widehat{p}_n(x)\right), \quad K_h(x) = \frac{1}{h^d} K\left(\frac{x}{h}\right),$$

where $\otimes$ denotes the convolution.

- We define $R_h = \text{Ridge}(p_h)$.

Smoothed Density Ridges

In particular, we focus on making inference for the smoothed version of the density, denoted as p_h:

$$p_h(x) = p \otimes K_h(x) = \mathbb{E}\left(\widehat{p}_n(x)\right), \quad K_h(x) = \frac{1}{h^d} K\left(\frac{x}{h}\right),$$

where $\otimes$ denotes the convolution.

- We define $R_h = \text{Ridge}(p_h)$.
- The advantages for focusing on R_h:
  - Always well-defined.
  - Topologically similar.
  - Asymptotically the same.
  - Fast rate of convergence.

Smoothed Density Ridges

In particular, we focus on making inference for the smoothed version of the density, denoted as p_h:

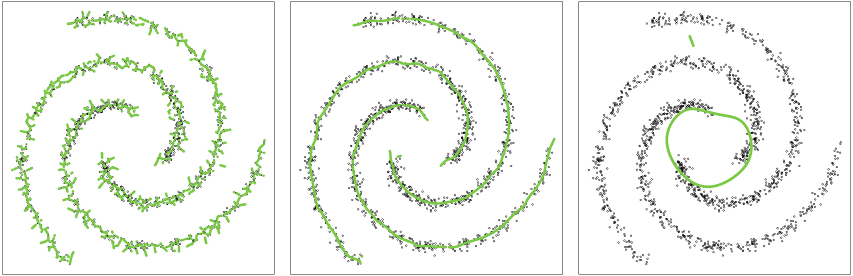
$$p_h(x) = p \otimes K_h(x) = \mathbb{E}\left(\widehat{p}_n(x)\right), \quad K_h(x) = \frac{1}{h^d} K\left(\frac{x}{h}\right),$$

where $\otimes$ denotes the convolution.

- We define $R_h = \text{Ridge}(p_h)$.
- The advantages for focusing on R_h:
  - Always well-defined.
  - Topologically similar.
  - Asymptotically the same.
  - Fast rate of convergence.
- One can always slightly undersmooth so that inference for R_h is asymptotically valid for R.

Bandwidth Selection for Density Ridges

Effect of Smoothing Bandwidth



Risk for Ridges

Let R and $\widehat{R}_n$ be the density ridges and their estimators.
Let

$$U_R \sim Unif\left(R\right), \quad U_{\widehat{R}_n} \sim Unif\left(\widehat{R}_n\right).$$

Risk for Ridges

Let R and $\widehat{R}_n$ be the density ridges and their estimators.
Let

$$U_R \sim Unif\left(R\right), \quad U_{\widehat{R}_n} \sim Unif\left(\widehat{R}_n\right).$$

Define

$$W_n = d(U_R, \widehat{R}_n), \quad \widetilde{W}_n = d(U_{\widehat{R}_n}, R)$$

be the projected distance of U_R onto $\widehat{R}_n$ and $U_{\widehat{R}_n}$ onto R.
We define L_2 risk as

$$\mathrm{Risk}_{2,n} = \frac{1}{2}\mathbb{E}(W_n^2 + \widetilde{W}_n^2).$$

Risk for Ridges

Let R and $\widehat{R}_n$ be the density ridges and their estimators.
Let

$$U_R \sim Unif\,(R)\,, \quad U_{\widehat{R}_n} \sim Unif\left(\widehat{R}_n\right).$$

Define

$$W_n = d(U_R, \widehat{R}_n), \quad \widetilde{W}_n = d(U_{\widehat{R}_n}, R)$$

be the projected distance of U_R onto $\widehat{R}_n$ and $U_{\widehat{R}_n}$ onto R.
We define L_2 risk as

$$\text{Risk}_{2,n} = \frac{1}{2}\mathbb{E}(W_n^2 + \widetilde{W}_n^2).$$

- This is a generalized mean integrated square errors.
- Similarly, one can define $\text{Risk}_{1,n}$ using L_1 loss.

Estimating Risks

We can use bootstrap or data splitting to estimate the risk $\mathrm{Risk}_{2,n}$. Let $\widehat{R}_n^*$ be the bootstrap version of $\widehat{R}_n$. Let

$$W_n^* = d(U_{\widehat{R}_n}, \widehat{R}_n^*), \quad \widetilde{W}_n^* = d(U_{\widehat{R}_n^*}, \widehat{R}_n)$$

Define

$$\widehat{\mathrm{Risk}}_{2,n} = \frac{1}{2}\mathbb{E}(W_n^{*2} + \widetilde{W}_n^{*2} | X_1, \cdots, X_n).$$

Estimating Risks

We can use bootstrap or data splitting to estimate the risk $\mathsf{Risk}_{2,n}$.
Let $\widehat{R}_n^*$ be the bootstrap version of $\widehat{R}_n$. Let

$$W_n^* = d(U_{\widehat{R}_n}, \widehat{R}_n^*), \quad \widetilde{W}_n^* = d(U_{\widehat{R}_n^*}, \widehat{R}_n)$$

Define

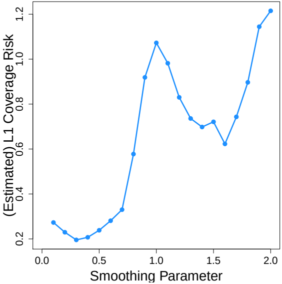
$$\widehat{\mathsf{Risk}}_{2,n} = \frac{1}{2}\mathbb{E}(W_n^{*2} + \widetilde{W}_n^{*2} | X_1, \cdots, X_n).$$

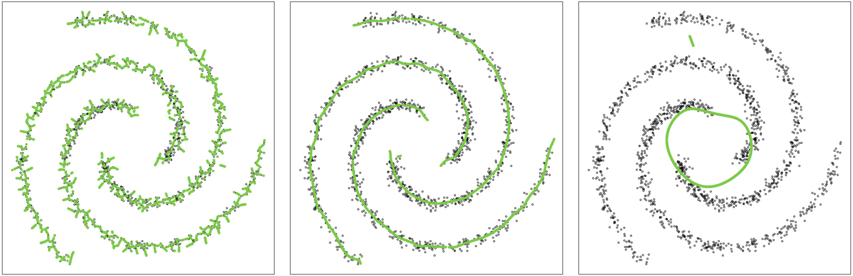
Theorem

Under regularity conditions,

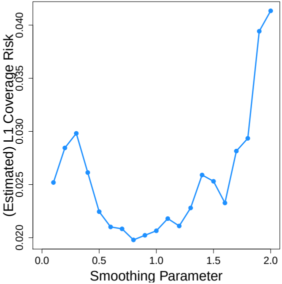
$$\frac{\widehat{\mathsf{Risk}}_{2,n}}{\mathsf{Risk}_{2,n}} \overset{P}{\to} 1, \quad \frac{\widehat{\mathsf{Risk}}_{1,n}}{\mathsf{Risk}_{1,n}} \overset{P}{\to} 1.$$

Bandwidth Selection via Risk Minimization





Application to Cosmology Dataset



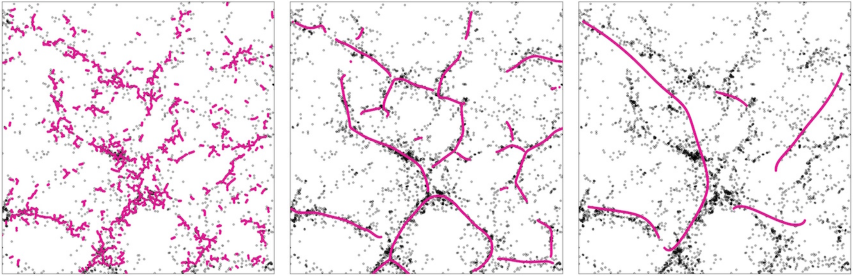
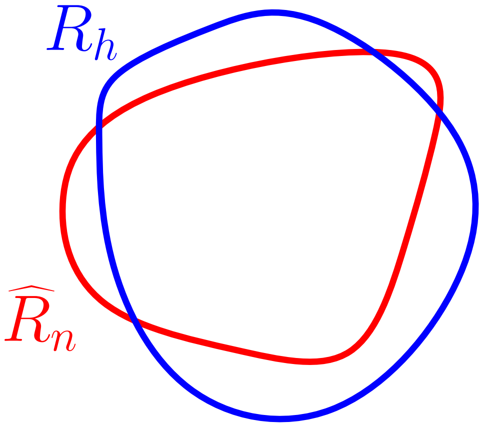
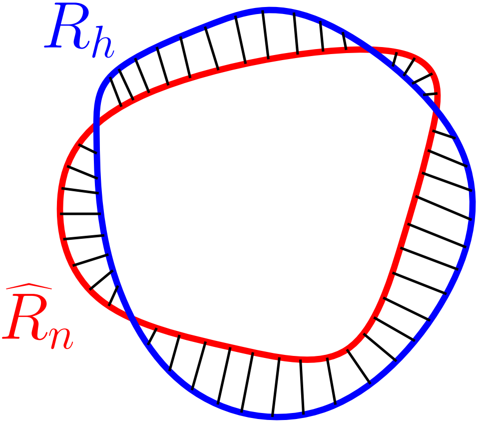


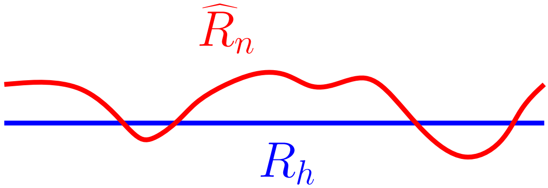
Illustration for Asymptotic Theory


R_h
$\widehat{R}_n$

Asymptotic Theory

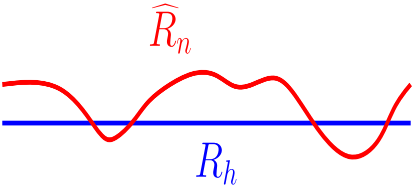


Asymptotic Theory



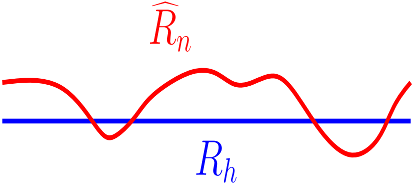
Asymptotic Theory

1. Thus, the projection distance $\approx$ a stochastic process.



Asymptotic Theory

1. Thus, the projection distance $\approx$ a stochastic process.
2. This stochastic process $\approx$ empirical process.



Asymptotic Theory

1. Thus, the projection distance $\approx$ a stochastic process.
2. This stochastic process $\approx$ empirical process.
3. $\text{Haus}(\widehat{D}_n, D_h) = \sup\{\text{projection distance}\} \approx \sup\{\text{Empirical process}\}$.