Interaction of abstract and concrete questions for kernel estimators

Tarn Duong

Laboratoire de Statistique Théorique et Appliquée

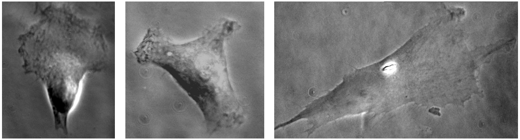
27 Nov 2012

Outline

1. Density estimation – Cellular compartments distributions
2. Derivative estimation – Sub-populations in mixed cell populations
3. Variable importance – Biomarker selection for Alzheimer's disease diagnosis

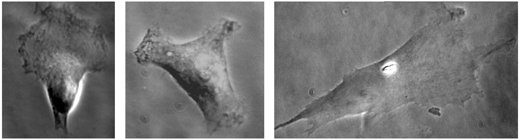
Application (1): Mammalian cells under a microscope

- Unconstrained mammalian cells grow into various shapes, making comparative analysis of large numbers of cells difficult

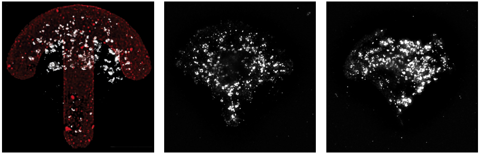


Application (1): Mammalian cells under a microscope

- Unconstrained mammalian cells grow into various shapes, making comparative analysis of large numbers of cells difficult

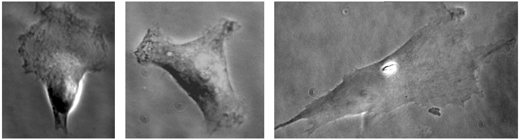


- Micro-patterns reproducibly induce cells to grow into standard shapes to facilitate comparisons

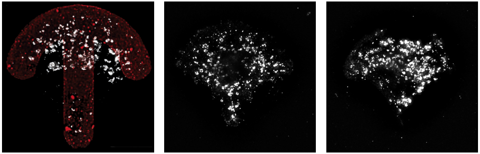


Application (1): Mammalian cells under a microscope

- Unconstrained mammalian cells grow into various shapes, making comparative analysis of large numbers of cells difficult



- Micro-patterns reproducibly induce cells to grow into standard shapes to facilitate comparisons



- Q: What is the density corresponding these point clouds? (Schauer et al., *Nature Meth.*, 2010)

Data smoothing

- Convert point clouds to density via *kernel density estimators*

$n = 200$
structures

$$X_1, \ldots, X_{200} \sim f$$

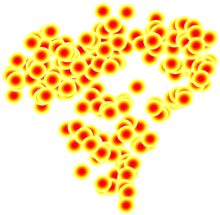
Data smoothing

- Convert point clouds to density via *kernel density estimators*

$n = 200$
structures

$X_1, \ldots, X_{200} \sim f$

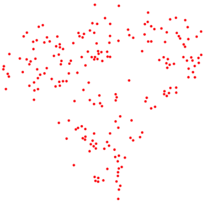
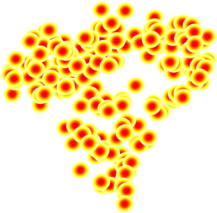
Kernels
(Convolution)



$K_{\mathbf{H}}(\boldsymbol{x} - \boldsymbol{X}_1), \ldots, K_{\mathbf{H}}(\boldsymbol{x} - \boldsymbol{X}_{200})$

Data smoothing

- Convert point clouds to density via *kernel density estimators*

$n = 200$ structures	Kernels (Convolution)	Kernel density estimator
		
$\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{200} \sim f$	$K_{\mathbf{H}}(\boldsymbol{x} - \boldsymbol{X}_1), \ldots, K_{\mathbf{H}}(\boldsymbol{x} - \boldsymbol{X}_{200})$	$\hat{f}_{\mathbf{H}}(\boldsymbol{x}) = \frac{1}{200} \sum_{i=1}^{200} K_{\mathbf{H}}(\boldsymbol{x} - \boldsymbol{X}_i)$

Data smoothing

- Convert point clouds to density via *kernel density estimators*

$n = 10397$ (35 cells) structures

$$X_1, \ldots, X_n \sim f$$

Kernel density estimator

$$\hat{f}_{\mathbf{H}}(x) = n^{-1} \sum_{i=1}^{n} K_{\mathbf{H}}(x - X_i)$$

Smoothing parameter estimation

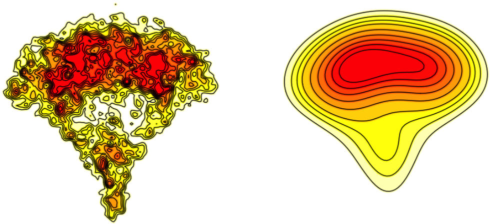
- Estimating smoothing parameter (bandwidth) matrix is crucial

Undersmoothed

$$\begin{bmatrix} 31.4 & 0.7 \\ 0.7 & 27.4 \end{bmatrix}$$

Oversmoothed

$$\begin{bmatrix} 3135.4 & 27.5 \\ 27.5 & 2737.5 \end{bmatrix}$$



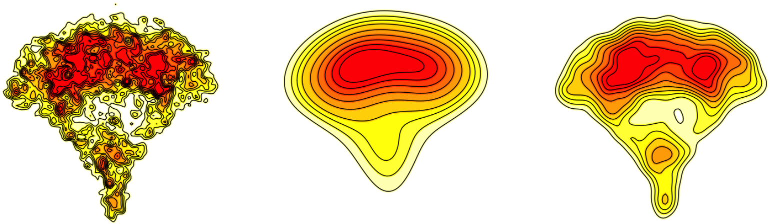
Smoothing parameter estimation

- Estimating smoothing parameter (bandwidth) matrix is crucial

Undersmoothed $\begin{bmatrix} 31.4 & 0.7 \\ 0.7 & 27.4 \end{bmatrix}$

Oversmoothed $\begin{bmatrix} 3135.4 & 27.5 \\ 27.5 & 2737.5 \end{bmatrix}$

Optimally smoothed $\begin{bmatrix} 313.5 & 2.7 \\ 2.7 & 273.7 \end{bmatrix}$



- Optimal smoothing parameter is minimiser of $\int [\hat{f}_{\mathbf{H}}(\boldsymbol{x}) - f(\boldsymbol{x})]^2 \, d\boldsymbol{x}$
- (Duong & Hazelton, *J. Nonparametr. Stat.*, 2003)

Optimal data-based bandwidth selection

- Optimality criterion: $\mathrm{MISE}(\mathbf{H}) = \int_{\mathbb{R}^d} \mathbb{E}[\hat{f}_{\mathbf{H}}(\boldsymbol{x}) - f(\boldsymbol{x})]^2 \, d\boldsymbol{x}$
- Under suitable regularity conditions, a Taylor's series expansion gives $\mathrm{MISE}(\mathbf{H}) = \mathrm{AMISE}(\mathbf{H})[1 + o(1)]$ where

$$\mathrm{AMISE}(\mathbf{H}) = n^{-1}|\mathbf{H}|^{-1/2}R(K) + \frac{1}{4}m_2(K)^2 \int_{\mathbb{R}^d} \mathrm{tr}^2(\mathbf{H}D^2 f(\boldsymbol{x}))\, d\boldsymbol{x}$$

and

$$R(K) = \int_{\mathbb{R}^d} K(\boldsymbol{x})^2 \, d\boldsymbol{x}$$
$$m_2(K)\mathbf{I}_d = \int_{\mathbb{R}^d} \boldsymbol{x}\boldsymbol{x}^T K(\boldsymbol{x})\, d\boldsymbol{x}$$
$$D^2 f(\boldsymbol{x}) = \text{Hessian matrix of second order partial derivatives of } f(\boldsymbol{x})$$

- Optimal selector is $\mathbf{H}_0 = \mathrm{argmin}_{\mathbf{H}\in\mathcal{F}} \mathrm{AMISE}(\mathbf{H})$ where $\mathcal{F}$ is the space of all symmetric, positive definite $d \times d$ matrices.
- Optimal data-based selector is $\hat{\mathbf{H}} = \mathrm{argmin}_{\mathbf{H}\in\mathcal{F}} \widehat{\mathrm{AMISE}}(\mathbf{H})$

Relative convergence rates

- Since $\mathbf{H}$ is a matrix, suitable definition of $\hat{\mathbf{H}}$ tends to $\mathbf{H}_0$ with relative rate $n^{-\alpha}, \alpha > 0$ if

$$\operatorname{vec}(\hat{\mathbf{H}} - \mathbf{H}_0) = O_p(n^{-\alpha}\mathbf{J}_{d^2}) \operatorname{vec} \mathbf{H}_0$$

where vec is the vector operator

$$\operatorname{vec}\begin{bmatrix} a_1 & a_4 & a_7 \\ a_2 & a_5 & a_8 \\ a_3 & a_6 & a_9 \end{bmatrix} = \begin{bmatrix} a_1 & a_2 & a_3 & a_4 & a_5 & a_6 & a_7 & a_8 & a_9 \end{bmatrix}^T$$

and $\mathbf{J}_d$ is the $d \times d$ matrix of all ones

Relative convergence rates

- Since $\mathbf{H}$ is a matrix, suitable definition of $\hat{\mathbf{H}}$ tends to $\mathbf{H}_0$ with relative rate $n^{-\alpha}, \alpha > 0$ if

$$\operatorname{vec}(\hat{\mathbf{H}} - \mathbf{H}_0) = O_p(n^{-\alpha}\mathbf{J}_{d^2}) \operatorname{vec}\mathbf{H}_0$$

  where vec is the vector operator

$$\operatorname{vec}\begin{bmatrix} a_1 & a_4 & a_7 \\ a_2 & a_5 & a_8 \\ a_3 & a_6 & a_9 \end{bmatrix} = \begin{bmatrix} a_1 & a_2 & a_3 & a_4 & a_5 & a_6 & a_7 & a_8 & a_9 \end{bmatrix}^T$$

  and $\mathbf{J}_d$ is the $d \times d$ matrix of all ones
- Direct computation of this rate is difficult, so indirect method is if $\mathbb{E}[\operatorname{vec}^T(\hat{\mathbf{H}} - \mathbf{H}_0)\operatorname{vec}(\hat{\mathbf{H}} - \mathbf{H}_0)] = O(n^{-2\alpha})$ implies that rate is $O(n^{-\alpha})$.
- Define $D_{\mathbf{H}} = \partial/(\partial \operatorname{vec}\mathbf{H})$, $D^2_{\mathbf{H}} = \partial/[(\partial \operatorname{vec}\mathbf{H})(\partial \operatorname{vec}^T \mathbf{H})]$ then

$$\begin{aligned}
\widehat{\mathrm{AMISE}}(\hat{\mathbf{H}}) &= (\widehat{\mathrm{AMISE}} - \mathrm{AMISE})(\hat{\mathbf{H}}) + \mathrm{AMISE}(\hat{\mathbf{H}}) \\
D_{\mathbf{H}}\widehat{\mathrm{AMISE}}(\hat{\mathbf{H}}) &= D_{\mathbf{H}}(\widehat{\mathrm{AMISE}} - \mathrm{AMISE})(\hat{\mathbf{H}}) + D_{\mathbf{H}}\mathrm{AMISE}(\hat{\mathbf{H}}) \\
&\sim \mathrm{D}_{\mathbf{H}}(\widehat{\mathrm{AMISE}} - \mathrm{AMISE})(\mathbf{H}_0) + D_{\mathbf{H}}\mathrm{AMISE}(\mathbf{H}_0) + \operatorname{vec}(\hat{\mathbf{H}} - \mathbf{H}_0) D^2_{\mathbf{H}}\mathrm{AMISE}(\mathbf{H}_0) \\
\operatorname{vec}(\hat{\mathbf{H}} - \mathbf{H}_0) &\sim [D^2_{\mathbf{H}}\mathrm{AMISE}(\mathbf{H}_0)]^{-1}\mathrm{D}_{\mathbf{H}}(\widehat{\mathrm{AMISE}} - \mathrm{AMISE})(\mathbf{H}_0)
\end{aligned}$$

- (Duong & Hazelton, *J. Multivar. Anal.*, 2005)

Density functional estimation (1)

- Requires estimation of $\int_{\mathbb{R}^d} \text{tr}^2(\mathbf{H}D^2 f(\boldsymbol{x}))\, d\boldsymbol{x}$
- Straightforward plugin estimator $\int_{\mathbb{R}^d} \text{tr}^2(\mathbf{H}D^2 \hat{f}_{\mathbf{H}}(\boldsymbol{x}))\, d\boldsymbol{x}$ suffers from two disadvantages
    1. Explicit numerical integration
    2. Bandwidth $\mathbf{H}$ which is optimal for estimating f is NOT optimal for calculating $D^2 f$

Density functional estimation (1)

- Requires estimation of $\int_{\mathbb{R}^d} \mathrm{tr}^2(\mathbf{H}D^2 f(\boldsymbol{x}))\, d\boldsymbol{x}$
- Straightforward plugin estimator $\int_{\mathbb{R}^d} \mathrm{tr}^2(\mathbf{H}D^2 \hat{f}_{\mathbf{H}}(\boldsymbol{x}))\, d\boldsymbol{x}$ suffers from two disadvantages
  1. Explicit numerical integration
  2. Bandwidth $\mathbf{H}$ which is optimal for estimating f is NOT optimal for calculating $D^2 f$
- Use matrix algebra/analysis to remove the requirement for numerical integration
- Let $\mathsf{D} = \left[\frac{\partial}{\partial x_1} \quad \cdots \quad \frac{\partial}{\partial x_d}\right]$ be the differential operator
- Define differential operator multiplication as $\dfrac{\partial}{\partial x_i}\dfrac{\partial}{\partial x_j} = \dfrac{\partial^2}{\partial x_i \partial x_j}$
- Hessian operator $D^2 = \mathsf{D}\,\mathsf{D}^T$

Density functional estimation (1)

- Requires estimation of $\int_{\mathbb{R}^d} \operatorname{tr}^2(\mathbf{H} D^2 f(\boldsymbol{x}))\, d\boldsymbol{x}$
- Straightforward plugin estimator $\int_{\mathbb{R}^d} \operatorname{tr}^2(\mathbf{H} D^2 \hat{f}_{\mathbf{H}}(\boldsymbol{x}))\, d\boldsymbol{x}$ suffers from two disadvantages
  1. Explicit numerical integration
  2. Bandwidth $\mathbf{H}$ which is optimal for estimating f is NOT optimal for calculating $D^2 f$
- Use matrix algebra/analysis to remove the requirement for numerical integration
- Let $\mathsf{D} = \begin{bmatrix} \frac{\partial}{\partial x_1} & \cdots & \frac{\partial}{\partial x_d} \end{bmatrix}$ be the differential operator
- Define differential operator multiplication as $\dfrac{\partial}{\partial x_i}\dfrac{\partial}{\partial x_j} = \dfrac{\partial^2}{\partial x_i \partial x_j}$
- Hessian operator $D^2 = \mathsf{D}\,\mathsf{D}^T$
- Let $\mathbf{A}, \mathbf{B}$ be $d \times d$ matrices, then $\operatorname{tr}(\mathbf{AB}) = \operatorname{tr}(\mathbf{BA}) = \operatorname{vec}^T(\mathbf{A}^T)\operatorname{vec}\mathbf{B}$
- Let $\boldsymbol{x}$ be d-vector then $\operatorname{vec}(\boldsymbol{x}\boldsymbol{x}^T) = (\boldsymbol{x} \otimes \boldsymbol{x})$ where $\otimes$ is the Kronecker (or tensor) product

Density functional estimation (2)

- Integrand $\mathrm{tr}(\mathbf{H}D^2 f(\boldsymbol{x})) = \mathrm{tr}(\mathbf{H}\,\mathsf{D}\,\mathsf{D}^T f(\boldsymbol{x})) = (\mathrm{vec}^T\,\mathbf{H})\,\mathrm{vec}(\mathsf{D}\,\mathsf{D}^T f(\boldsymbol{x})) = (\mathrm{vec}^T\,\mathbf{H})((\mathsf{D} \otimes \mathsf{D}) f(\boldsymbol{x}))$

Density functional estimation (2)

- Integrand $\operatorname{tr}(\mathbf{H}D^2 f(\boldsymbol{x})) = \operatorname{tr}(\mathbf{H}\,\mathsf{D}\,\mathsf{D}^T f(\boldsymbol{x})) = (\operatorname{vec}^T \mathbf{H}) \operatorname{vec}(\mathsf{D}\,\mathsf{D}^T f(\boldsymbol{x})) = (\operatorname{vec}^T \mathbf{H})((\mathsf{D} \otimes \mathsf{D}) f(\boldsymbol{x}))$
- Integrand

$$\begin{aligned}
\operatorname{tr}^2(\mathbf{H}D^2 f(\boldsymbol{x})) &= (\operatorname{vec}^T \mathbf{H})((\mathsf{D} \otimes \mathsf{D}) f(\boldsymbol{x}))((\mathsf{D} \otimes \mathsf{D})^T f(\boldsymbol{x}))(\operatorname{vec} \mathbf{H}) \\
&= \operatorname{tr}\left[(\operatorname{vec} \mathbf{H})(\operatorname{vec}^T \mathbf{H})((\mathsf{D} \otimes \mathsf{D}) f(\boldsymbol{x}))((\mathsf{D} \otimes \mathsf{D})^T f(\boldsymbol{x}))\right] \\
&= (\operatorname{vec}^T \mathbf{H} \otimes \operatorname{vec}^T \mathbf{H})((\mathsf{D} \otimes \mathsf{D}) f(\boldsymbol{x})) \otimes ((\mathsf{D} \otimes \mathsf{D}) f(\boldsymbol{x}))
\end{aligned}$$

- Last line decouples role of $\mathbf{H}$ and f

Density functional estimation (2)

- Integrand $\operatorname{tr}(\mathbf{H}D^2 f(\boldsymbol{x})) = \operatorname{tr}(\mathbf{H}\,\mathsf{D}\,\mathsf{D}^T f(\boldsymbol{x})) = (\operatorname{vec}^T \mathbf{H}) \operatorname{vec}(\mathsf{D}\,\mathsf{D}^T f(\boldsymbol{x})) = (\operatorname{vec}^T \mathbf{H})((\mathsf{D} \otimes \mathsf{D}) f(\boldsymbol{x}))$
- Integrand

$$\begin{aligned}
\operatorname{tr}^2(\mathbf{H}D^2 f(\boldsymbol{x})) &= (\operatorname{vec}^T \mathbf{H})((\mathsf{D} \otimes \mathsf{D}) f(\boldsymbol{x}))((\mathsf{D} \otimes \mathsf{D})^T f(\boldsymbol{x}))(\operatorname{vec} \mathbf{H}) \\
&= \operatorname{tr}\left[(\operatorname{vec} \mathbf{H})(\operatorname{vec}^T \mathbf{H})((\mathsf{D} \otimes \mathsf{D}) f(\boldsymbol{x}))((\mathsf{D} \otimes \mathsf{D})^T f(\boldsymbol{x}))\right] \\
&= (\operatorname{vec}^T \mathbf{H} \otimes \operatorname{vec}^T \mathbf{H})((\mathsf{D} \otimes \mathsf{D}) f(\boldsymbol{x})) \otimes ((\mathsf{D} \otimes \mathsf{D}) f(\boldsymbol{x}))
\end{aligned}$$

- Last line decouples role of $\mathbf{H}$ and f
- Integral

$$\begin{aligned}
\int_{\mathbb{R}^d} \operatorname{tr}^2(\mathbf{H}D^2 f(\boldsymbol{x}))\, d\boldsymbol{x} &= (\operatorname{vec}^T \mathbf{H} \otimes \operatorname{vec}^T \mathbf{H}) \left[\int_{\mathbb{R}^d} ((\mathsf{D} \otimes \mathsf{D}) f(\boldsymbol{x})) \otimes ((\mathsf{D} \otimes \mathsf{D}) f(\boldsymbol{x}))\, d\boldsymbol{x} \right] \\
&= (\operatorname{vec}^T \mathbf{H} \otimes \operatorname{vec}^T \mathbf{H}) \left[\int_{\mathbb{R}^d} (\mathsf{D} \otimes \mathsf{D} \otimes \mathsf{D} \otimes \mathsf{D}) f(\boldsymbol{x}) f(\boldsymbol{x})\, d\boldsymbol{x} \right]
\end{aligned}$$

Density functional estimation (2)

- Integrand $\mathrm{tr}(\mathbf{H}D^2 f(\boldsymbol{x})) = \mathrm{tr}(\mathbf{H}\,\mathsf{D}\,\mathsf{D}^T f(\boldsymbol{x})) = (\mathrm{vec}^T\,\mathbf{H})\,\mathrm{vec}(\mathsf{D}\,\mathsf{D}^T f(\boldsymbol{x})) = (\mathrm{vec}^T\,\mathbf{H})((\mathsf{D}\otimes\mathsf{D})f(\boldsymbol{x}))$
- Integrand

$$\begin{aligned}\mathrm{tr}^2(\mathbf{H}D^2 f(\boldsymbol{x})) &= (\mathrm{vec}^T\,\mathbf{H})((\mathsf{D}\otimes\mathsf{D})f(\boldsymbol{x}))((\mathsf{D}\otimes\mathsf{D})^T f(\boldsymbol{x}))(\mathrm{vec}\,\mathbf{H})\\ &= \mathrm{tr}\left[(\mathrm{vec}\,\mathbf{H})(\mathrm{vec}^T\,\mathbf{H})((\mathsf{D}\otimes\mathsf{D})f(\boldsymbol{x}))((\mathsf{D}\otimes\mathsf{D})^T f(\boldsymbol{x}))\right]\\ &= (\mathrm{vec}^T\,\mathbf{H}\otimes\mathrm{vec}^T\,\mathbf{H})((\mathsf{D}\otimes\mathsf{D})f(\boldsymbol{x}))\otimes((\mathsf{D}\otimes\mathsf{D})f(\boldsymbol{x}))\end{aligned}$$

- Last line decouples role of $\mathbf{H}$ and f
- Integral

$$\begin{aligned}\int_{\mathbb{R}^d}\mathrm{tr}^2(\mathbf{H}D^2 f(\boldsymbol{x}))\,d\boldsymbol{x} &= (\mathrm{vec}^T\,\mathbf{H}\otimes\mathrm{vec}^T\,\mathbf{H})\left[\int_{\mathbb{R}^d}((\mathsf{D}\otimes\mathsf{D})f(\boldsymbol{x}))\otimes((\mathsf{D}\otimes\mathsf{D})f(\boldsymbol{x}))\,d\boldsymbol{x}\right]\\ &= (\mathrm{vec}^T\,\mathbf{H}\otimes\mathrm{vec}^T\,\mathbf{H})\left[\int_{\mathbb{R}^d}(\mathsf{D}\otimes\mathsf{D}\otimes\mathsf{D}\otimes\mathsf{D})f(\boldsymbol{x})f(\boldsymbol{x})\,d\boldsymbol{x}\right]\end{aligned}$$

- Let $\boldsymbol{\psi}_4 = \int_{\mathbb{R}^d}\mathsf{D}^{\otimes 4} f(\boldsymbol{x})f(\boldsymbol{x})\,d\boldsymbol{x} = \mathbb{E}[\mathsf{D}^{\otimes 4} f(\boldsymbol{X})]$ since $\boldsymbol{X}\sim f$
- Usual kernel estimator is $\hat{\boldsymbol{\psi}}_4(\mathbf{G}) = n^{-1}\sum_{i=1}^{n}\mathsf{D}^{\otimes 4}\hat{f}_{\mathbf{G}}(\boldsymbol{X}_i) = n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathsf{D}^{\otimes 4}K_{\mathbf{G}}(\boldsymbol{X}_i-\boldsymbol{X}_j)$

  where $\mathbf{G}$ is a pilot bandwidth, independent of $\mathbf{H}$
- (Chacón & Duong, *Test*, 2010)

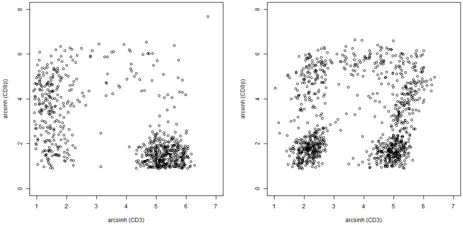
Plug-in bandwidth selection

- $\widehat{\text{AMISE}}(\mathbf{H}) = n^{-1}|\mathbf{H}|^{-1/2}R(K) + \frac{1}{4}m_2(K)^2(\text{vec}\,\mathbf{H} \otimes \text{vec}\,\mathbf{H})^T\hat{\boldsymbol{\psi}}_4(\mathbf{G})$
- $\hat{\mathbf{H}}$ converges to $\mathbf{H}_0$ with rate $n^{-2/(d+6)}$

Application (2): Sub-populations in mixed cell populations

- Flow cytometer machine measures the fluorescence of cells as a proxy for their properties
- Q: Is there a significantly different sub-population between a control and diseased patient?

Control Graft-versus-host



Higher order Taylor's expansion

- Q: What is the Taylor's expansion of a multivariate function $f : \mathbb{R}^d \to \mathbb{R}$?
- A:

$$f(\boldsymbol{x}) = f(\boldsymbol{x}_0) + (\boldsymbol{x} - \boldsymbol{x}_0)^T Df(\boldsymbol{x}_0) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_0)^T D^2 f(\boldsymbol{x}_0)(\boldsymbol{x} - \boldsymbol{x}_0) + \dots$$

Higher order Taylor's expansion

- Q: What is the Taylor's expansion of a multivariate function $f : \mathbb{R}^d \to \mathbb{R}$?
- A:

$$
\begin{aligned}
f(\boldsymbol{x}) &= f(\boldsymbol{x}_0) + (\boldsymbol{x} - \boldsymbol{x}_0)^T Df(\boldsymbol{x}_0) + \tfrac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_0)^T D^2 f(\boldsymbol{x}_0)(\boldsymbol{x} - \boldsymbol{x}_0) + \ldots \\
&= f(\boldsymbol{x}_0) + (\boldsymbol{x} - \boldsymbol{x}_0)^T \, \mathsf{D} f(\boldsymbol{x}_0) + \tfrac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_0)^{T \otimes 2} \, \mathsf{D}^{\otimes 2} f(\boldsymbol{x}_0) + \ldots
\end{aligned}
$$

Higher order Taylor's expansion

- Q: What is the Taylor's expansion of a multivariate function $f : \mathbb{R}^d \to \mathbb{R}$?
- A:

$$
\begin{aligned}
f(\boldsymbol{x}) &= f(\boldsymbol{x}_0) + (\boldsymbol{x} - \boldsymbol{x}_0)^T Df(\boldsymbol{x}_0) + \tfrac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_0)^T D^2 f(\boldsymbol{x}_0)(\boldsymbol{x} - \boldsymbol{x}_0) + \ldots \\
&= f(\boldsymbol{x}_0) + (\boldsymbol{x} - \boldsymbol{x}_0)^T \mathsf{D} f(\boldsymbol{x}_0) + \tfrac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_0)^{T\otimes 2} \mathsf{D}^{\otimes 2} f(\boldsymbol{x}_0) + \ldots \\
&\quad + \tfrac{1}{j!}(\boldsymbol{x} - \boldsymbol{x}_0)^{T\otimes j} \mathsf{D}^{\otimes j} f(\boldsymbol{x}_0) + \ldots
\end{aligned}
$$

Higher order Taylor's expansion

- Q: What is the Taylor's expansion of a multivariate function $f : \mathbb{R}^d \to \mathbb{R}$?
- A:

$$\begin{aligned} f(\boldsymbol{x}) &= f(\boldsymbol{x}_0) + (\boldsymbol{x}-\boldsymbol{x}_0)^T Df(\boldsymbol{x}_0) + \tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{x}_0)^T D^2 f(\boldsymbol{x}_0)(\boldsymbol{x}-\boldsymbol{x}_0) + \dots \\ &= f(\boldsymbol{x}_0) + (\boldsymbol{x}-\boldsymbol{x}_0)^T \, \mathsf{D} f(\boldsymbol{x}_0) + \tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{x}_0)^{T\otimes 2} \, \mathsf{D}^{\otimes 2} f(\boldsymbol{x}_0) + \dots \\ &\quad + \tfrac{1}{j!}(\boldsymbol{x}-\boldsymbol{x}_0)^{T\otimes j} \, \mathsf{D}^{\otimes j} f(\boldsymbol{x}_0) + \dots \end{aligned}$$

- Q: What is the Taylor's expansion of a vector valued function $\boldsymbol{f} : \mathbb{R}^d \to \mathbb{R}^p$?

Higher order Taylor's expansion

- Q: What is the Taylor's expansion of a multivariate function $f : \mathbb{R}^d \to \mathbb{R}$?
- A:

$$\begin{aligned} f(\boldsymbol{x}) &= f(\boldsymbol{x}_0) + (\boldsymbol{x}-\boldsymbol{x}_0)^T Df(\boldsymbol{x}_0) + \tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{x}_0)^T D^2 f(\boldsymbol{x}_0)(\boldsymbol{x}-\boldsymbol{x}_0) + \dots \\ &= f(\boldsymbol{x}_0) + (\boldsymbol{x}-\boldsymbol{x}_0)^T \,\mathsf{D} f(\boldsymbol{x}_0) + \tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{x}_0)^{T\otimes 2}\,\mathsf{D}^{\otimes 2} f(\boldsymbol{x}_0) + \dots \\ &\quad + \tfrac{1}{j!}(\boldsymbol{x}-\boldsymbol{x}_0)^{T\otimes j}\,\mathsf{D}^{\otimes j} f(\boldsymbol{x}_0) + \dots \end{aligned}$$

- Q: What is the Taylor's expansion of a vector valued function $\boldsymbol{f} : \mathbb{R}^d \to \mathbb{R}^p$?
- A:

$$\begin{aligned} \boldsymbol{f}(\boldsymbol{x}) &= \boldsymbol{f}(\boldsymbol{x}_0) + [\mathbf{I}_p \otimes (\boldsymbol{x}-\boldsymbol{x}_0)^T]\,\mathsf{D} \boldsymbol{f}(\boldsymbol{x}_0) + \tfrac{1}{2}[\mathbf{I}_p \otimes (\boldsymbol{x}-\boldsymbol{x}_0)^{T\otimes 2}]\,\mathsf{D}^{\otimes 2} \boldsymbol{f}(\boldsymbol{x}_0) + \dots \\ &\quad + \tfrac{1}{j!}[\mathbf{I}_p \otimes (\boldsymbol{x}-\boldsymbol{x}_0)^{T\otimes j}]\,\mathsf{D}^{\otimes j} \boldsymbol{f}(\boldsymbol{x}_0) + \dots \end{aligned}$$

- (Chacón, Duong & Wand, *Stat. Sinica*, 2011)

Kernel density derivative estimation

- Derivatives contain information that is not available in the density itself (e.g. local extrema)
- Generalise to estimation of derivatives of density function f

Kernel density derivative estimation

- Derivatives contain information that is not available in the density itself (e.g. local extrema)
- Generalise to estimation of derivatives of density function f
- Kernel estimator of r-th derivative $\mathsf{D}^{\otimes r} f$ is

$$\mathsf{D}^{\otimes r} \hat{f}_{\mathbf{H}}(\boldsymbol{x}) = n^{-1} \sum_{i=1}^{n} \mathsf{D}^{\otimes r} K_{\mathbf{H}}(\boldsymbol{x} - \boldsymbol{X}_i)$$

- L_2 error is

$$\begin{aligned}\mathrm{AMISE}(\mathsf{D}^{\otimes r} \hat{f}_{\mathbf{H}}) &= n^{-1}|\mathbf{H}|^{-1/2} \operatorname{tr}((\mathbf{H}^{-1})^{\otimes r}\mathbf{R}(\mathsf{D}^{\otimes r} K)) \\ &\quad + \tfrac{1}{4} m_2(K)^2 \operatorname{tr}\left[(\mathbf{I}_{d^r} \otimes \operatorname{vec}\mathbf{H} \operatorname{vec}^T \mathbf{H})\mathbf{R}(\mathsf{D}^{\otimes (r+2)} f)\right]\end{aligned}$$

  where $\mathbf{R}(\boldsymbol{g}) = \int \boldsymbol{g}(\boldsymbol{x})\boldsymbol{g}(\boldsymbol{x})^T \, d\boldsymbol{x}$

- Analogously $\mathbf{H}_0 = \operatorname{argmin}_{\mathbf{H}\in\mathcal{F}} \mathrm{AMISE}(\mathsf{D}^{\otimes r} \hat{f}_{\mathbf{H}})$ and $\hat{\mathbf{H}} = \operatorname{argmin}_{\mathbf{H}\in\mathcal{F}} \widehat{\mathrm{AMISE}}(\mathsf{D}^{\otimes r} \hat{f}_{\mathbf{H}})$
- (Chacón & Duong , *Elec. J. Stat*, 2012?)

Local modal regions

- Local mode of f is $\{\boldsymbol{x} : Df(\boldsymbol{x}) = 0, D^2f(\boldsymbol{x}) < 0\}$
- Local modal region is $\{\boldsymbol{x} : D^2f(\boldsymbol{x}) < \epsilon_2\}$

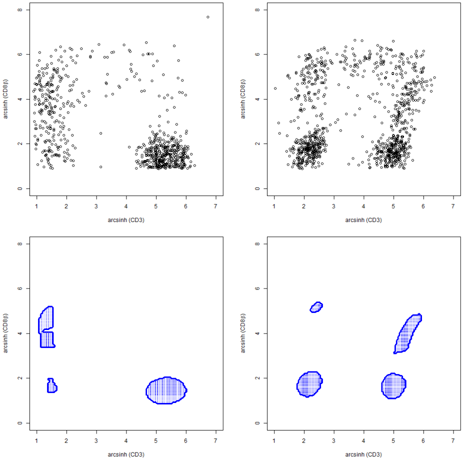
Local modal regions

- Local mode of f is $\{\boldsymbol{x} : Df(\boldsymbol{x}) = 0, D^2 f(\boldsymbol{x}) < 0\}$
- Local modal region is $\{\boldsymbol{x} : D^2 f(\boldsymbol{x}) < \epsilon_2\}$
- Rejection region of local hypothesis tests $H_0(\boldsymbol{x}) : \|D^2 f(\boldsymbol{x})\| = 0$
- Test statistic $W(\boldsymbol{x}) = \|\mathbf{S}(\boldsymbol{x})^{-1/2}\, \mathsf{D}^{\otimes 2} \hat{f}_{\mathbf{H}}(\boldsymbol{x})\|_2^2 \mathrel{\dot\sim} \chi_d^2$ where $\mathbf{S}(\boldsymbol{x})$ is an estimator of $\operatorname{Var} \mathsf{D}^{\otimes 2} \hat{f}_{\mathbf{H}}(\boldsymbol{x})$
- (Duong et al., *Comp. Stat. Data. Anal.*, 2008)

Application (2): Sub-populations in mixed cell populations

- Q: Is there a significantly different sub-population between a control and diseased patient?

Control　　Graft-versus-host



Kernel variable importance (1)

- Q: What the most important biomarkers (variables) for diagnosing Alzheimer's disease in patients?
- Let $\boldsymbol{X} = (X_1, \dots, X_d)$ be d variables collected from a patient
- Let $f_{X_k,1}$ be the (marginal) density of the k-th variable control patients, $f_{k,2}$ for Alzheimer's patients, f_{X_k} for the pooled control and Alzheimer's patients
- Let Z be indicator that patient has AD, $Z \sim \text{Bern}(\pi)$

Kernel variable importance (1)

- Q: What the most important biomarkers (variables) for diagnosing Alzheimer's disease in patients?
- Let $\boldsymbol{X} = (X_1, \ldots, X_d)$ be d variables collected from a patient
- Let $f_{X_k,1}$ be the (marginal) density of the k-th variable control patients, $f_{k,2}$ for Alzheimer's patients, f_{X_k} for the pooled control and Alzheimer's patients
- Let Z be indicator that patient has AD, $Z \sim \text{Bern}(\pi)$
- Mutual information for the k-th variable

$$
\begin{aligned}
Q_k &= \int_{\mathbb{R}^2} [f_{X_k,Z}(x,z) - f_{X_k}(x) f_Z(z)]^2 \, dxdz \\
&= \int_{\mathbb{R}^2} [f_{X_k|Z}(x) f_Z(z) - f_{X_k}(x) f_Z(z)]^2 \, dxdz \\
&= \sum_{j=1}^{2} [\mathbb{P}(Z = z_j)]^2 \int_{\mathbb{R}} [f_{X_k,j}(x) - f_{X_k}(x)]^2 \, dx
\end{aligned}
$$

- Large values of mutual information imply important variable to discriminate between groups

Kernel variable importance (2)

- Components of mutual information $Q_{k,1} = \pi^2(\psi + \psi_1 - 2\psi_1')$ where

$$\psi = \int_{\mathbb{R}} f(x)^2\, dx = \mathbb{E} f_{X_k}(X_k), X_k \sim f_{X_k}$$

$$\psi_1 = \int_{\mathbb{R}} f_{X_k,1}(x)^2\, dx = \mathbb{E} f_{X_k,1}(X), X_{k,1} \sim f_{X_k,1}$$

$$\psi_1' = \int_{\mathbb{R}} f_{X_k,1}(x) f_{X_k}(x)\, dx = \mathbb{E} f_{X_k,1}(X), X_k \sim f_{X_k}$$

can be estimated using usual kernel-based U-statistics

Kernel variable n-tuple importance

- Mutual information for pair of (k_1, k_2)-th variables

$$Q_{k_1,k_2} = \pi^2 \int_{\mathbb{R}^2} [f_{X_{k_1},X_{k_2},1}(x_1,x_2) - f_{X_{k_1},X_{k_2}}(x_1,x_2)]^2 \, dx_1 dx_2 + (1-\pi)^2 \int_{\mathbb{R}^2} [f_{X_{k_1},X_{k_2},2}(x_1,x_2) - f_{X_{k_1},X_{k_2}}(x_1,x_2)]^2 \, dx_1 dx_2$$

- Select n-tuples of variables can discover higher order interactions missed by serial selection of single variables