Various ways to summarize a curve with a number: estimation and applications

Cristina Butucea, Paris-Est Marne-la-Vallée

October 14, 2010

- - Claims amount during one year $X_1, \ldots, X_n$ are randomly distributed.
  - Costs of various incidents in the enterprise.
  Density model: $X_1, \ldots, X_n$ i.i.d., belonging to $\mathbb{R}^d$, $d \geq 1$, having distribution function F, density function f.

- - Claims amount during one year $X_1, \ldots, X_n$ are randomly distributed.
  - Costs of various incidents in the enterprise.
  Density model: $X_1, \ldots, X_n$ i.i.d., belonging to $\mathbb{R}^d$, $d \geq 1$, having distribution function F, density function f.
- - Explain how the risk of a portfolio is related to the market risk.
  Regression model: $(X_1, Y_1), \ldots, (X_n, Y_n)$ i.i.d.,
  $Y_i = m(X_i) + \xi_i$, ξ_i are centered, with finite variance.
  First fit a distribution to the regressors X_i, then estimate $m : [0, 1] \to \mathbb{R}$.

- - Black-Scholes model:

$$\frac{dS_t}{S_t} = \mu dt + \sigma dW_t.$$

If we fix the time interval $[0, T]$ and change $\mu(t)$ and $\varepsilon\sigma(t)$, for small ε: **Gaussian white model**
- Capital of an insurance company, when the time is rescaled is assumed to follow such a model.

- - Black-Scholes model:

$$\frac{dS_t}{S_t} = \mu dt + \sigma dW_t.$$

If we fix the time interval $[0, T]$ and change $\mu(t)$ and $\varepsilon\sigma(t)$, for small ε: **Gaussian white model**
- Capital of an insurance company, when the time is rescaled is assumed to follow such a model.

- Spectral density in time-series, etc.

- **Parametric model** (claim amount is distributed according to an exponential distribution - estimate its parameter).
  + good estimators, good rates and asymptotic properties;
  + simple to work out

- **Parametric model** (claim amount is distributed according to an exponential distribution - estimate its parameter).
  + good estimators, good rates and asymptotic properties;
  + simple to work out
- vs. **Nonparametric model** (claim amount is a function, 3 times continuously differentiable)
  + more realistic in many applications
  + choice of methods
  - worse rates (for same sample size)
  - curse of dimensionality ($d \geq 3$).

- **Parametric model** (claim amount is distributed according to an exponential distribution - estimate its parameter).
  + good estimators, good rates and asymptotic properties;
  + simple to work out
- vs. **Nonparametric model** (claim amount is a function, 3 times continuously differentiable)
  + more realistic in many applications
  + choice of methods
  - worse rates (for same sample size)
  - curse of dimensionality ($d \geq 3$).
- **Idea**: sometimes we need far less than to recover the whole function!

Nonparametric estimation:
1) smoothness hypothesis,
2) choice of method - kernel estimator, projection on orthogonal series, on wavelets (MPEG);
3) tuning of method's parameters according to the assumed smoothness. (bandwidth, number of estimated coefficients, etc.)

Example: best rate (means squared error) for estimating a $\mathcal{C}^s$ function from a sample of size n is $n^{-s/(2s+d)}$.
This is worse than the parametric rate $1/\sqrt{n}$, higher is the dimension d.
So, estimate only functionals, with better rates, when less information is needed out of data.

- 1) for fixed x_0, value at the point: $f(x_0)$.
(the volatility at a certain time value)

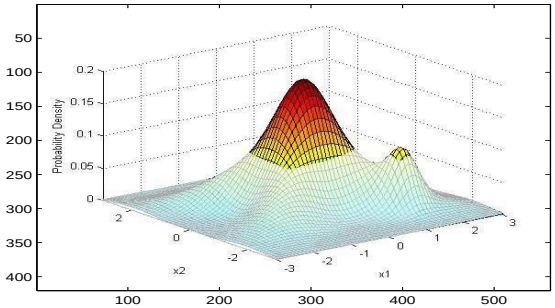
- 1) for fixed x_0, value at the point: $f(x_0)$.
  (the volatility at a certain time value)
- 2) for fixed $a < b$: $\int_a^b f(x)dx$.
  (given the VaR of last year, what is the risk this year at that value: $P(S > VaR)$?)

- 1) for fixed x_0, value at the point: $f(x_0)$.
  (the volatility at a certain time value)
- 2) for fixed $a < b$: $\int_a^b f(x)dx$.
  (given the VaR of last year, what is the risk this year at that value: $P(S > VaR)$?)
- 3) $I = \int f^2(x)dx$ (goodness-of-fit tests)

- 1) for fixed x_0, value at the point: $f(x_0)$.
  (the volatility at a certain time value)
- 2) for fixed $a < b$: $\int_a^b f(x)dx$.
  (given the VaR of last year, what is the risk this year at that value: $P(S > VaR)$?)
- 3) $I = \int f^2(x)dx$ (goodness-of-fit tests)
- 4) entropy of a distribution $Ent = \int_{f(x)>0} f(x)\log(1/f(x))dx$
  (measures the 'disorder in the randomness').
  $Ent(N(0,\sigma^2) = \log(\sigma\sqrt{2\pi}) + 1/2$.

- 1) for fixed x_0, value at the point: $f(x_0)$.
  (the volatility at a certain time value)
- 2) for fixed $a < b$: $\int_a^b f(x)dx$.
  (given the VaR of last year, what is the risk this year at that value: $P(S > VaR)$?)
- 3) $I = \int f^2(x)dx$ (goodness-of-fit tests)
- 4) entropy of a distribution $Ent = \int_{f(x)>0} f(x)\log(1/f(x))dx$
  (measures the 'disorder in the randomness').
  $Ent(N(0,\sigma^2) = \log(\sigma\sqrt{2\pi}) + 1/2$.
- 5) excess mass at level $\nu > 0$: $E(\nu) = \int (f(x) - \nu)_+ dx$.

- 1) for fixed x_0, value at the point: $f(x_0)$.
  (the volatility at a certain time value)
- 2) for fixed $a < b$: $\int_a^b f(x)dx$.
  (given the VaR of last year, what is the risk this year at that value: $P(S > VaR)$?)
- 3) $I = \int f^2(x)dx$ (goodness-of-fit tests)
- 4) entropy of a distribution $Ent = \int_{f(x)>0} f(x)\log(1/f(x))dx$
  (measures the 'disorder in the randomness').
  $Ent(N(0,\sigma^2) = \log(\sigma\sqrt{2\pi}) + 1/2$.
- 5) excess mass at level $\nu > 0$: $E(\nu) = \int (f(x) - \nu)_+ dx$.
- Note that some are integrated functionals: $\int \Phi(f(x))dx$,
  where
  3) $\Phi(u) = u^2$; 4) $\Phi(u) = u\log(1/u)$, for $u > 0$;
  5) $\Phi(u) = (|u| - \nu)_+$.



Detect clustering in a population of claim amounts, for example.
How many different subpopulations in my data?

Fitting a distribution to the data
Given f_0, and a sample of size n, test

$$\begin{aligned} H_0 &\quad f \equiv f_0 \\ H_1 &\quad d(f, f_0) \geq C\phi_n, \end{aligned}$$

where $\phi_n \searrow 0$ with n is the testing rate.
If we write $\int (f(x) - f_0(x))^2 dx \geq C^2\phi_n^2$, the test statistic is an estimator of

$$\int (f(x) - f_0(x))^2 dx = \int f^2 - 2\int f_0 f + \int f_0^2$$

.
It is sufficient to estimate $\int f^2$.

Regression model:
Given m_0, and a sample of size n, test

$$\begin{aligned} H_0 &\quad m \equiv m_0 \\ H_1 &\quad d(m, m_0) \geq C\phi_n, \end{aligned}$$

where $\phi_n \searrow 0$ with n is the testing rate.
We may take $d(m, m_0) = \int_a^b (m - m_0)$ and test statistic attains parametric rates of estimation.

or $d(m, m_0) = \int |m - m_0|$ which is related to the excess mass for density functions.

Given two independent samples of i.i.d. observations having unknown probability densities f and g, decide whether they come from the same population or not:

$$\begin{array}{ll} H_0 & f \equiv g \\ H_1 & d(f,g) \geq C\phi_n, \end{array}$$

where $\phi_n \searrow 0$ with n is the testing rate.
Again, we may have various $d(f,g)$ and corresponding test statistics with their own behaviors.

Estimate $\int \Phi(f(x))dx$.
If Φ is $\mathcal{C}^4$ smooth and f is $\mathcal{C}^s$ smooth, then the rate is

$$\begin{cases} n^{-1/2}, & \text{smoothness of } f : s \geq \frac{1}{4} \\ n^{-2s/(4s+d)}, & s < 1/4. \end{cases}$$

Bickel and Ritov ('88), Birgé and Massart ('95), Kerkyacharian and Picard ('96), Tribouley ('00).

They use Taylor expansion of Φ at 4th order and estimate $\int f^2$ and $\int f^3$. Actually, same rate as for $\int f^2$.

- If Φ (periodized) belongs to $\mathcal{C}^1$, $\mathcal{C}^2$: Nemirovski ('00).

- If Φ (periodized) belongs to $\mathcal{C}^1$, $\mathcal{C}^2$: Nemirovski ('00).
- For excess mass, Φ (periodized) belongs to $\mathcal{C}^0$.
  Most related to Lepski, Nemirovski and Spokoiny ('99).
  In particular, $\Phi(u) = |u|$, $\Phi \in \mathcal{C}^0$.

- If Φ (periodized) belongs to $\mathcal{C}^1$, $\mathcal{C}^2$: Nemirovski ('00).
- For excess mass, Φ (periodized) belongs to $\mathcal{C}^0$.
  Most related to Lepski, Nemirovski and Spokoiny ('99).
  In particular, $\Phi(u) = |u|$, $\Phi \in \mathcal{C}^0$.
- B., Tribouley, Mougeot (2007) generalize to density model, $d \geq 1$, Besov classes of density functions (uniformly bounded away from 0).

First step

- Use Fourier expansion of $\Phi(u) = (|u| - \nu)_+$, instead of Taylor expansion.
  For any Φ such that its periodized form is in $\mathcal{C}^0$,

$$c_k := \int_{-1}^{1} \Phi(u) \cos(\pi k u) du \sim \frac{1}{k^2}, \text{ for large } k.$$

First step

- Use Fourier expansion of $\Phi(u) = (|u| - \nu)_+$, instead of Taylor expansion.
  For any Φ such that its periodized form is in $\mathcal{C}^0$,

$$c_k := \int_{-1}^{1} \Phi(u) \cos(\pi k u) du \sim \frac{1}{k^2}, \text{ for large } k.$$

- As Φ is symmetric, then

$$\Phi(u) \sim \sum_{k=0}^{N} c_k(\nu) \cos(\pi k f(t)) dt.$$

2nd step

- We need to estimate $\cos(\pi k f(t))$ for all $t \in K$.

Funny thing

If $\varepsilon \sim \mathcal{N}(0, \lambda^2)$, *then*

$$e^{\pi^2 k^2 \lambda^2 / 2} \mathbb{E}[\cos(\pi k (f(t) + \varepsilon))] = \cos(\pi k f(t)).$$

2nd step

- We need to estimate $\cos(\pi k f(t))$ for all $t \in K$.

Funny thing

If $\varepsilon \sim \mathcal{N}(0, \lambda^2)$, *then*

$$e^{\pi^2 k^2 \lambda^2/2} \mathbb{E}[\cos(\pi k(f(t) + \varepsilon))] = \cos(\pi k f(t)).$$

- The trouble here is: λ^2 is the variance of a density estimator, thus
  -the density estimator has not a Gaussian law
  -the variance is proportional to unknown $f(t)$.

- We estimate $f(t)$, $t \in K$ by a wavelet estimator $f_j(t)$ and its variance $\lambda_j^2(t)$ by $\hat{\lambda}_j^2(t)$.

- We estimate $f(t)$, $t \in K$ by a wavelet estimator $f_j(t)$ and its variance $\lambda_j^2(t)$ by $\hat{\lambda}_j^2(t)$.
- Finally,

$$\hat{E}(\nu) = \sum_{k=0}^{N} c_k(\nu) \int_K e^{\frac{\pi^2 k^2}{2} \min\left\{\hat{\lambda}_j^2(t), \frac{\gamma 2^{jd}}{n}\right\}} \cos(\pi k \hat{f}_j(t)) dt.$$

- We estimate $f(t)$, $t \in K$ by a wavelet estimator $f_j(t)$ and its variance $\lambda_j^2(t)$ by $\hat{\lambda}_j^2(t)$.
- Finally,

$$\hat{E}(\nu) = \sum_{k=0}^{N} c_k(\nu) \int_K e^{\frac{\pi^2 k^2}{2} \min\left\{\hat{\lambda}_j^2(t), \frac{\gamma 2^{jd}}{n}\right\}} \cos(\pi k \hat{f}_j(t)) dt.$$

- Remark: algorithmically very easy to compute it simultaneously for different values of ν.

Rates

- Our procedure is thus shown to attain the rate

$$(n \log(n))^{-s/(2s+d)},$$

slightly faster than the pointwise rate for estimating f.

Rates

- Our procedure is thus shown to attain the rate

$$(n\log(n))^{-s/(2s+d)},$$

slightly faster than the pointwise rate for estimating f.

- Lower bounds for gaussian white noise and $d = 1$: Lepski, Nemirovski and Spokoiny ('99):

$$\frac{(n\log(n))^{-s/(2s+1)}}{\log(n)}$$

Cai and Low (Manuscript):

$$\frac{(n\log(n))^{-s/(2s+1)}}{\log(n)^{1/(2s+1)}}.$$

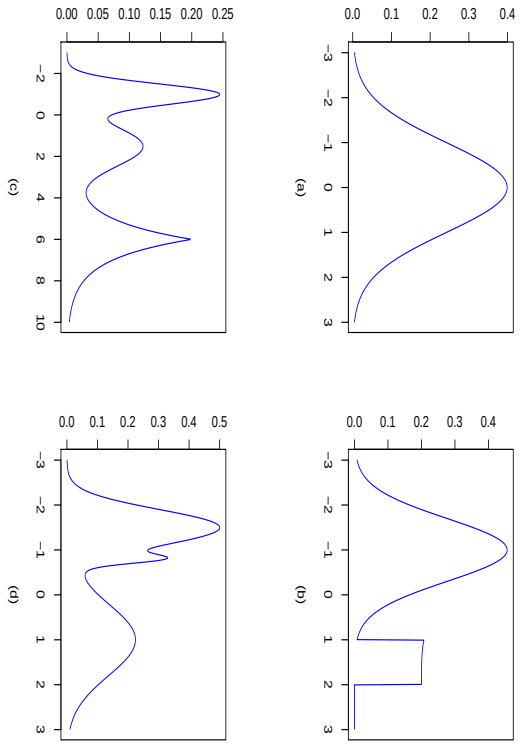
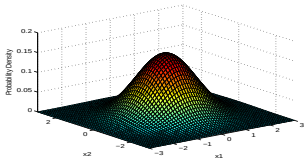


Figure: Set of studied densities. (a): standard gaussian; (b): mixture of gaussian and uniform; (c): mixture of 2 gaussian and laplace; (d):mixture of gaussian with isolated spoke.

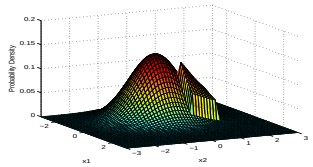
Univariate densities

f	n	$\widetilde{E}_2^{PI}$	$\widetilde{E}_2^*$	$\widetilde{E}_2^{PI}/\widetilde{E}_2^*$	$\widetilde{p}_2$	$\widetilde{E}_\infty^{PI}$	$\widetilde{E}_\infty^*$	$\widetilde{E}_\infty^{PI}/\widetilde{E}_\infty^*$	$\widetilde{p}_\infty$
a	100	0.00504	0.00542	0.93	0.45	0.04450	0.04855	0.92	0.45
a	1000	0.00079	0.00066	1.19	0.70	0.01937	0.01765	1.10	0.75
a	10000	0.00008	0.00006	1.32	0.55	0.00602	0.00590	1.02	0.60
b	100	0.00354	0.00533	0.66	0.20	0.03742	0.04989	0.75	0.30
b	1000	0.00147	0.00086	1.71	0.90	0.03217	0.02133	1.51	0.95
b	10000	0.00170	0.00083	2.06	1.00	0.03645	0.02445	1.49	1.00
c	100	0.00520	0.01027	0.51	0.15	0.04132	0.04924	0.84	0.30
c	1000	0.00077	0.00036	2.17	0.80	0.01745	0.01274	1.37	0.80
c	10000	0.00075	0.00021	3.64	1.00	0.01714	0.00938	1.83	1.00
d	100	0.03271	0.01857	1.76	1.00	0.11473	0.08293	1.38	1.00
d	1000	0.00975	0.00346	2.82	1.00	0.05985	0.03606	1.66	1.00
d	10000	0.00248	0.00063	3.91	1.00	0.02975	0.01525	1.95	1.00

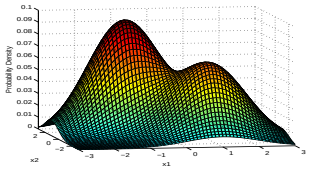
Table: Comparison of $\hat{E}^*$ and $\hat{E}^{PI}$ in mean integrated squared error and in mean error of the sup-norm, over $K = 20$ Monte-Carlo simulations, for various sizes of samples $n = 100, 1000, 10000$.



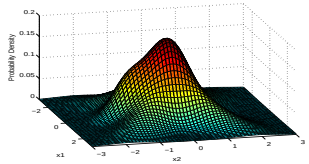
(A)



(B)



(C)



(D)

Figure: Set of studied 2D densities. (A): 2D Gaussian. (B): mixture of 2D gaussian and uniform. (C): mixture of two 2D gaussian. (D): Mixture of three 2D gaussian.

Bivariate densities

f	n	$\widetilde{E}_2^{PI}$	$\widetilde{E}_2^*$	$\widetilde{E}_2^{PI}/\widetilde{E}_2^*$	$\widetilde{p}_2$	$\widetilde{E}_\infty^{PI}$	$\widetilde{E}_\infty^*$	$\widetilde{E}_\infty/\widetilde{E}_\infty^*$	$\widetilde{p}_\infty$
A	400	0.01685	0.00870	1.94	0.95	0.07435	0.05675	1.31	0.95
A	1000	0.00948	0.00394	2.41	1.00	0.05445	0.03525	1.54	1.00
A	10000	0.00635	0.00263	2.41	1.00	0.04524	0.02867	1.58	1.00
B	400	0.10397	0.04628	2.25	1.00	0.17667	0.12818	1.38	1.00
B	1000	0.07460	0.02943	2.53	1.00	0.15398	0.10660	1.44	1.00
B	10000	0.04747	0.01985	2.39	1.00	0.12894	0.08901	1.45	1.00
C	400	0.01184	0.00555	2.13	1.00	0.06442	0.04609	1.40	1.00
C	1000	0.00906	0.00432	2.09	1.00	0.05462	0.03717	1.47	1.00
C	10000	0.00379	0.00134	2.83	1.00	0.03566	0.02081	1.71	1.00
D	400	0.26965	0.26187	1.03	0.55	0.34953	0.34350	1.02	0.55
D	1000	0.25655	0.24609	1.04	0.85	0.33525	0.32628	1.03	0.85
D	10000	0.23705	0.22277	1.06	1.00	0.31686	0.30497	1.04	1.00

Table: Comparison of $\hat{E}^*$ and $\hat{E}^{PI}$ in mean integrated squared error and in mean error of the sup-norm, over $K = 20$ Monte-Carlo simulations, for various sizes of samples $n = 400, 1000, 10000$.