Spatially Smoothed Kernel Density Estimation via Generalized Empirical Likelihood

Kuangyu Wen & Ximing Wu

Texas A&M University

Info-Metrics Institute Conference: Recent Innovations in Info-Metrics
October 2014
Motivation

- Nonparametric density estimation offers a flexible alternative to parametric method.
- To obtain the same level of precision as that of its parametric counterpart, a larger sample size is desired in nonparametric estimation due to its slower convergence rate.
- Given similarity/dependence among distributions across geographic units, efficiency gain is possible via information pooling; particularly so in nonparametric estimations with small sample size for each unit.
- Example: Crop yield distributions are typically estimated (individually) at the county level. Taking into account their spatial dependence may improve estimation.
Nonparametric estimation of spatially dependent densities

- An information-theoretic nonparametric estimation procedure
  - Weighted Kernel Density Estimator (WKDE) that satisfies moment conditions implied by spatial dependence among units
  - Observation weights are the implied Generalized Empirical Likelihood (GEL) probabilities associated with spatial moment conditions.

- GEL offers a flexible yet simple framework to incorporate out-of-sample information into density estimation.

- Computationally inexpensive; joint estimation of a large number of spatially dependent densities is avoided.
Talk plan

- Weighted KDE (WKDE) via the GEL
- Spatial smoothing of KDE
- Monte Carlo Simulations
- Empirical example
Kernel Density Estimation (KDE)

- Data: $X_1, \ldots, X_T$ i.i.d. from a continuously differentiable distribution $F$ with density $f$

- KDE:

$$
\hat{f}(x) = \frac{1}{Th} \sum_{t=1}^{T} K \left( \frac{x - X_t}{h} \right) \equiv \frac{1}{T} \sum_{t=1}^{T} K_h(x - X_t)
$$

where $K$ is a kernel function and $h$ the bandwidth.

- Weighted KDE (WKDE)

$$
\hat{f}(x) = \sum_{t=1}^{T} p_t K_h(x - X_t)
$$

where $p = (p_1, \ldots, p_T)$ are non-negative observation weights that sum to one.
Incorporation of out-of-sample information

- Incorporation of out-of-sample information, for instance population moments or theoretical properties, may enhance the KDE.
- Suppose out-of-sample information comes in the form
  \[ E[g_m(x)] = \mu_m, \ m = 1, \ldots, M, \]
  where \( g_m \) are known real functions.
- Chen (1997) uses WKDE to incorporate out-of-sample information; the weights is calculated via empirical likelihood (Owen, 1988; 2001)

\[
\begin{align*}
\max_{p_1, \ldots, p_T} & \sum_{t=1}^{T} \log p_t \\
\text{s.t.} & \sum_{t=1}^{T} p_t g_m(X_t) = \mu_m, \ m = 1, \ldots, M, \\
& p_t \geq 0, \sum_{t=1}^{T} p_t = 1.
\end{align*}
\]
Empirical likelihood weighted KDE:

\[ \hat{f}_{EL}(x) = \sum_{t=1}^{T} \hat{p}_t K_h(x - X_t) \]

where \( \hat{p} = (\hat{p}_1, \ldots, \hat{p}_T) \) are the implied EL probabilities.

Under mild regularity conditions

\[
\text{abias}(\hat{f}_{EL}(x)) = \text{abias}(\hat{f}(x)) + o(n^{-1})
\]
\[
\text{avar}(\hat{f}_{EL}(x)) = \text{avar}(\hat{f}(x)) - r(x)f^2(x)/n + o(n^{-1})
\]

where \( r(x) : \mathbb{R} \rightarrow \mathbb{R}^+ \) depends on the unknown density \( f \) and moment functions \( g = (g_1, \ldots, g_M) \).

An important implication: \( \hat{f}_{EL} \) can use the optimal bandwidth for \( \hat{f} \); joint search of \( p \) and bandwidth is not required.
Extension to Generalized Empirical Likelihood

- Cressie-Read power discrepancy

$$CR_\nu = \frac{1}{\nu(\nu + 1)} \sum_{t=1}^{T} ((np_t)^{-\nu} - 1)$$

- Generalized Empirical Likelihood (GEL)

$$\min_{p_1, \ldots, p_T} \frac{1}{\nu(\nu + 1)} \sum_{t=1}^{T} ((np_t)^{-\nu} - 1)$$

s.t. $$\sum_{t=1}^{T} p_t g_m(X_t) = \mu_m, m = 1, \ldots, M,$$

$$p_t \geq 0, \sum_{t=1}^{T} p_t = 1.$$ 

- $$\nu = 0 \implies EL$$
- $$\nu = 1 \implies \text{Exponential Tilting (ET) or maximum entropy}$$
Spatial smoothing of KDE

- Consider $N$ spatially dependent densities $f_i, i = 1, \ldots, N$; for each $i$, we observe $Y_i = (Y_{i,1}, \ldots, Y_{i,T})$ i.i.d. from $f_i$.
- Given selected moment functions $g = (g_1, \ldots, g_M)$, define their sample analog for each unit

$$\hat{g}(i) = \frac{1}{T} \sum_{t=1}^{T} g(Y_{i,t}), i = 1, \ldots, N.$$ 

- Spatial weights for unit $i$: $w_i = \{w_{i,j}\}_{j=1}^{N}$; $w_{i,j} \geq 0$ and $\sum_{j=1}^{N} w_{i,j} = 1$.
- Spatially smoothed sample moments

$$\tilde{g}(i) = \sum_{j=1}^{N} w_{i,j} \left( \frac{1}{T} \sum_{t=1}^{T} g(Y_{j,t}) \right) = \sum_{j=1}^{N} w_{i,j} \hat{g}(j)$$
- Estimate \( f_i \) using the WKDE \( \hat{f}_{EL} \) or \( \hat{f}_{ET} \) subject to spatial moment conditions \( \tilde{g}(i), \ i = 1, \ldots, N \).

- For implication, we need to specify
  - spatial weights \( w = \{ w_1, \ldots, w_N \} \)
  - moment functions \( g = (g_1, \ldots, g_M) \)
Construction of spatial weights

- $d_{i,j} \geq 0$: the spatial distance between unit $i$ and $j$
- $N^k_i$: an indicator set including unit $i$ and its $k$ nearest neighbors
- Spatial weights for unit $i$:

$$w_{i,j} = \frac{\exp\left(-d_{i,j}^2 / \gamma_i\right) \mathbb{1}(j \in N^k_i)}{\sum_{j=1}^N \exp\left(-d_{i,j}^2 / \gamma_i\right) \mathbb{1}(j \in N^k_i)}, j = 1, \ldots, N,$$

where $\mathbb{1}$ is the indicator function and

$$\gamma_i = \frac{1}{k} \sum_{j \in N^k_i} d_{i,j}$$
Consider for now the Exponential Tilting estimator $\hat{f}_{\text{ET}}$. It can be shown to be asymptotically equivalent to a minimum cross entropy density estimator with the KDE $\hat{f}$ as the reference density:

$$
\tilde{f}_{\text{ET}} = \arg \min_f \int f(x) \log \frac{f(x)}{\hat{f}(x)} \, dx
$$

s.t. $$
\int f(x) \, dx = 1,
$$
$$
\int g_m(x) f(x) \, dx = \mu_m, m = 1, \ldots, M.
$$

The solution takes the form

$$
\tilde{f}_{\text{ET}}(x) = \hat{f}(x) \exp \left\{ \tilde{\lambda}_0 + \sum_{m=1}^{M} \tilde{\lambda}_m g_m(x) \right\} = \hat{f}_{\text{ET}}(x) + o_p(n^{-1})
$$

where $\tilde{\lambda}_0$ is a normalization constant.
Thus $\hat{f}_{ET}$ is seen to modify $\hat{f}$ with a multiplicative adjustment, which consists of a basis expansion in exponent. Therefore the selection of moment functions for $\hat{f}_{ET}$ amounts to the selection of basis functions for nonparametric minimal cross entropy density estimation.

Power series, $g(x) = (x, x^2, x^3, ...)$, is customarily used, but can be sensitive to outliers and require the existence of population moments up to order $M$.

We use spline basis due to its flexibility and robustness.; e.g.,

$$g(x) = (x, x^2, (x - K_1)^2_+, (x - K_2)^2_+, \ldots, (x - K_M)^2_+),$$

where $(x)_+ = \max(0, x)$ and $K_1, \ldots, K_M$ are spline knots. Only the first two moments are required to exist.

The resultant density can be viewed as a hybrid of KDE and log spline estimator (Kooperberg and Stone, 1992; Stone, et al., 1997)
Monte Carlo Simulations

- $N = 100$ locations in a unit square $\mathcal{I}$
- Construct a $20 \times 20$ grid evenly spaced over $\mathcal{I}$
- Randomly draw 100 points from the grid without replacement
- Calculate the distance matrix $\mathbf{D} = [d_{i,j}]$
Assume $f_i \sim \text{Skewed Normal} \left( \mu_i, \sigma_i, \alpha_i \right)$ for $i = 1, \cdots, N$.

- $\mu = (\mu_1, \cdots, \mu_N)^\top$, $\sigma = (\sigma_1, \cdots, \sigma_N)^\top$ and $\alpha = (\alpha_1, \cdots, \alpha_N)^\top$.
- A $N$-dimensional vector $\theta$ follows spatial error model $\text{SpE} (\beta, \lambda, \sigma_\epsilon)$ if
  \[
  \theta = \beta + \eta \\
  \eta = \lambda W \eta + \epsilon,
  \]
  where $\epsilon \sim \mathcal{N} (0, \sigma_\epsilon^2 I_N)$ and the spatial weighting matrix $W$ is constructed as
  \begin{enumerate}
  \item $W^* = [w_{i,j}^*]$, $w_{i,j}^* = 1$ if $i \neq j$ and the location $i$ is one of the 5 nearest neighbors of the location $j$; otherwise $w_{i,j}^* = 0$;
  \item row standardization, i.e. $W = [w_{i,j}]$, $w_{i,j} = w_{i,j}^* / \sum_{j=1}^{N} w_{i,j}^*$.
  \end{enumerate}
We set $\mu \sim \text{SpE}(-1, 0.8, 0.03)$, $\log(\sigma) \sim \text{SpE}(0, 0.8, 0.03)$ and $\alpha \sim \text{SpE}(-0.75, 0.8, 0.03)$.

Each vector of shape parameter is given by

$$\theta = \beta + (I_N - \lambda W)^{-1} \epsilon.$$

Randomly draw $\epsilon$ to obtain a set of values for each $\mu$, $\sigma$ and $\alpha$ respectively.
Generate data $Y_i = \{Y_{i,t}\}_{t=1}^T$ from $f_i$, $i = 1, \ldots, N$

Sample size $T = 30$ and $60$

We consider the KDE $\hat{f}$ and two WKDEs: $\hat{f}_{EL}$ and $\hat{f}_{ET}$.

For each unit, KDE bandwidth is selected according to the rule of thumb.

Moment functions: quadratic splines with two knots at the 33th and 66th percentiles of the pooled sample:

$$g(y) = (y, y^2, (y - \bar{Q}_{33})^2_+, (y - \bar{Q}_{66})^2_+),$$

where $(y)_+ = \max(y, 0)$ and $\bar{Q}_\alpha$ is the $\alpha$th sample percentile.

Number of nearest neighbors in the calculation of spatial weights: $k = 5$

Number of repetitions: $R = 500$
Second experiment

- Construct $f_i$ as a two-component normal mixture, i.e.

$$f_i = \omega_i \phi(\mu_{1,i}, \sigma_{1,i}) + (1 - \omega_i) \phi(\mu_{2,i}, \sigma_{2,i}).$$

- Parameters: $\Phi^{-1}(\omega) \sim \text{SpE}(\Phi^{-1}(0.2), 0.6, 0.03)$, $\mu_1 \sim \text{SpE}(-3, 0.6, 0.2)$, $\mu_2 \sim \text{SpE}(0, 0.6, 0.1)$, $\log(\sigma_1) \sim \text{SpE}(\log(1.5), 0.6, 0.03)$ and $\log(\sigma_2) \sim \text{SpE}(0, 0.6, 0.03)$. 

![Graph of the mixture distribution](image)
Evaluation of performance

- Denote by $\hat{f}_i^{(r)}$ an estimate of $f_i$ in the $r$th experiment, $i = 1, \ldots, N$ and $r = 1, \ldots, R$.

- Global performance: average mean integrated squared error

$$
\frac{1}{NR} \sum_{i=1}^{N} \sum_{r=1}^{R} \int \left( \hat{f}_i^{(r)} - f_i \right)^2
$$

- Tail performance: let $Q_{i,5}$ be the 5% quantile of the $f_i$ and $\hat{\pi}_i^{(r)} = \int_{x \leq Q_{i,5}} \hat{f}_i^{(r)}(x)dx$, the average mean squared error of tail probability estimation

$$
\frac{1}{NR} \sum_{i=1}^{N} \sum_{r=1}^{R} \left( \hat{\pi}_i^{(r)} - 5\% \right)^2
$$
Simulation results

MSE of KDE and ratio of WKDE MSE relative to that of KDE

<table>
<thead>
<tr>
<th></th>
<th>KDE</th>
<th>WKDE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>EL</td>
</tr>
<tr>
<td>$T = 30$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SN</td>
<td>0.0176</td>
<td>63.51%</td>
</tr>
<tr>
<td></td>
<td>0.0017</td>
<td>56.08%</td>
</tr>
<tr>
<td>NM</td>
<td>0.0148</td>
<td>77.01%</td>
</tr>
<tr>
<td></td>
<td>0.0013</td>
<td>34.66%</td>
</tr>
<tr>
<td>$T = 60$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SN</td>
<td>0.0100</td>
<td>62.88%</td>
</tr>
<tr>
<td></td>
<td>0.0009</td>
<td>62.95%</td>
</tr>
<tr>
<td>NM</td>
<td>0.0088</td>
<td>75.49%</td>
</tr>
<tr>
<td></td>
<td>0.0006</td>
<td>37.91%</td>
</tr>
</tbody>
</table>