Analysis with Kernel Density Estimation

S. Gliske

University of Michigan / HERMES Collaboration

Transverse Parton Structure of the Hadron
Yerevan, Armenia
25 June, 2009

Outline

- Background and Motivation
  - Terminology
  - Kernel Density Estimation (KDEs)
- Example 1: Azimuthal Asymmetries with Small Statistics
  - Excl. ϕ-meson A_{UT}
- Example 2: Unfolding Acceptance/Smearing Effects
  - SIDIS π $\cos(n\phi)$
- Conclusion

Motivation and Background

Expensive Machines vs. Machine Learning

- Often we encounter the situation that an existing machine could measure additional observables if only. . .
- Common solution is to add new hardware component
- New hardware is not always feasible due to time/money constrains.
- Exist many Machine Learning techniques optimized to get the most information out of available data
- This talk comprises just one tool, KDEs, and two particularly challenging analysis: azimuthal moments with small statistics and unfolding radiative and detector smearing/acceptance.

Terminology

Density Estimation: The process of estimating $p(\boldsymbol{x})$ given $\left\{\boldsymbol{x}^{(i)}\right\}_{i=0}^{N} \sim p(x)$. Generally, one selects a model $\widehat{p}(\boldsymbol{x};\boldsymbol{\alpha})$ and determines $\widehat{\boldsymbol{\alpha}}$ to optimize $p(\boldsymbol{x}) \approx \widehat{p}(\boldsymbol{x};\widehat{\boldsymbol{\alpha}})$

Parameters: The parameters $\boldsymbol{\alpha}$ in the model.

Model Parameters: Distinct from $\boldsymbol{\alpha}$, these describe general features of the model.

Parametric Model: A model such that the number of parameters α_i is fixed.

Non-parametric Model: A model such that the number of parameters α_i is determined by the data.

- All hadronic structure analysis involves density estimation at some level.
- Histograms are discontinuous, parametric density estimators.
- Continuous, non-parametric estimators especially preferable in the case of
  - Small statistics
  - High dimension
  - Concerned about effect bin width/placement effects
- Also useful in classification problems

From histograms to KDEs

- Think of each bin of a histogram as a column of small boxes, one box per data point within the bin.
- Instead of aligning each box with the bin edges, center each box at the given data point $\boldsymbol{\mu}^{(i)}$.
- Rather than using boxes, a select a shape $K\left(\boldsymbol{x}-\boldsymbol{\mu}^{(i)}\right)$ (kernel function).
- Allow the scale of the kernel to vary per data point, $K\left(\left(H^{(i)}\right)^{-1}\left(\boldsymbol{x}-\boldsymbol{\mu}^{(i)}\right)\right)$.
- The result: a KDE

$$\widehat{p}(\boldsymbol{x}) = \frac{1}{N}\sum_{i=1}^{N} K\left(\left(H^{(i)}\right)^{-1}\left(\boldsymbol{x}-\boldsymbol{\mu}^{(i)}\right)\right). \tag{1}$$

- The matrix $H^{(i)}$ is the bandwidth matrix, and is usually chosen to be diagonal.
- Kernels assumed normalized and centered:

$$\int d^D\boldsymbol{u}\; K(\boldsymbol{u}) = 1, \quad \int d^D\boldsymbol{u}\; u_i K(\boldsymbol{u}) = 0. \tag{2}$$

Histogram vs KDE

Histograms	KDEs
Discontinuous	Continuous
Parametric (generally)	Non-parametric
Slower convergence.	Faster convergence
Must select bin widths, placement	Only select kernel shape
Several types of bias due to “bin effects”	Negligible bias due to shape
Fast to compute and evaluate	More computationally intensive

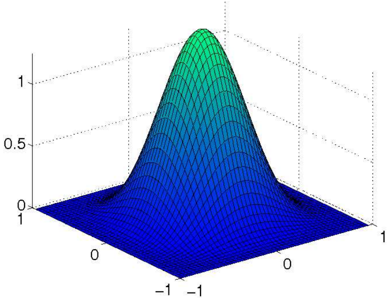
“It can be shown that, under weak assumptions, there cannot exist a non-parametric estimator that converges at a faster rate than the kernel estimator” —Wikipedia

- Primary Reference for KDEs: Silverman, B.W. (1986) “Density estimation: for statistics and data analysis.”
- KDEs still area of active research, esp. for high (> 2) dimensions

Further details

- Clara Kernel is designed for high-dimension, high-data KDEs

$$K\left(\left(H^{(i)}\right)^{-1}\left(\boldsymbol{x}-\boldsymbol{\mu}^{(i)}\right)\right) = \mathcal{N}^D \prod_{k=1}^{D}\left[1-\left(\frac{x_k-\mu_k^{(i)}}{h_k}\right)^2\right]^\gamma$$



Normalized Clara Kernel, $\gamma = 4$

- Cartesian yet approximately radially symmetric
- Evaluates in $O(D)$ time
- Bandwidths optimized by minimizing cross validation or hold-out-one estimates of *KL* divergence (reduces to maximum likelihood problem), using Simulated Annealing.
- Must choose model for how bandwidths vary with evaluation point
  - Good choice: piecewise constants according to decision tree structured domains

Fourier Moments of KDEs

- Integrals of 1D Clara Kernels with cosine and sine functions

$$\int d\phi^G \cos(n\phi^G) K_i^G(\phi_G) = \left(\frac{1}{2nh_i}\right)^{\gamma+1/2} \frac{\Gamma(2\gamma+2)\sqrt{\pi}}{\Gamma(\gamma+1)} J_{\gamma+1/2}(nh_i) \cos\left(n\mu^{(i)}\right); \quad (3)$$

$$\int d\phi^G \sin(n\phi^G) K_i^G(\phi_G) = \left(\frac{1}{2nh_i}\right)^{\gamma+1/2} \frac{\Gamma(2\gamma+2)\sqrt{\pi}}{\Gamma(\gamma+1)} J_{\gamma+1/2}(nh_i) \sin\left(n\mu^{(i)}\right). \quad (4)$$

- Assume data $\{\boldsymbol{x}^{(i)}\}_{i=1}^{N} \sim p(\boldsymbol{x})$ and KDE estimate $\widehat{p}(\boldsymbol{x}) = \sum_i K_i(\boldsymbol{x})$.
- Compare Monte Carlo Integral vs. Integral of KDE

$$2\langle\cos(n\phi)\rangle \approx \frac{1}{N}\sum_i \cos\left(n\phi^{(i)}\right), \quad (5)$$

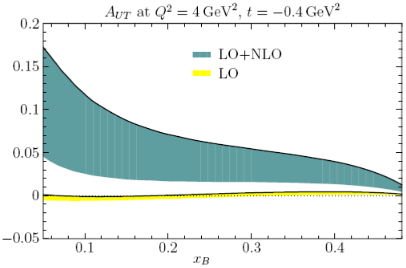
$$\approx \frac{1}{N}\sum_i \left(\frac{1}{2nh_i}\right)^{\gamma+1/2} \frac{\Gamma(2\gamma+2)\sqrt{\pi}}{\Gamma(\gamma+1)} J_{\gamma+1/2}(nh_i) \cos\left(n\phi^{(i)}\right). \quad (6)$$

- Equal only in limit $nh_i \to 0$, i.e. when kernel function becomes δ-function
- For $h_i > 0$, KDE moments smaller in magnitude–larger effect for larger n
- Similar effect for any Kernel function
- Indirect relation between KDE Fourier moment's accuracy and amount of data
- However, since bias is quantified, can correct for it in some circumstances

Example 1: Azimuthal Asymmetries

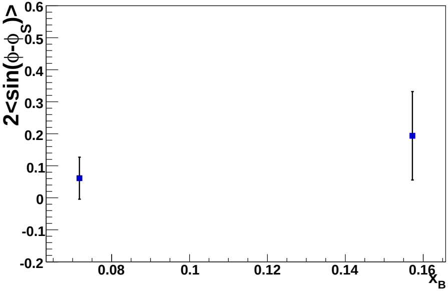
Example Details

- Consider exclusive ϕ lepto-production from polarized proton, $ep^{\uparrow} \rightarrow e'\phi p'$
- HERMES had about 500 events in 2002-2005
- Consider tuned PYTHIA Monte Carlo of about same size
- Consider studying whether any x_B dependence can be determined, to compare with Diehl/Kuglar model (arXiv:0708.1121v1)
- Difficult, as expected dependence is on the order of the statistical uncertainty

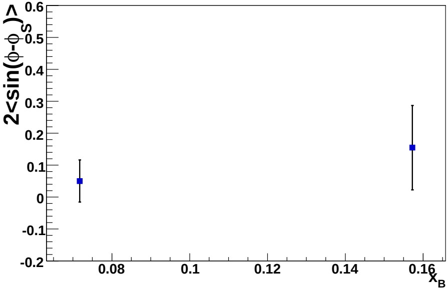


Diehl/Kuglar Model of $\sin(\phi - \phi_s)$ moment of the cross section, versus x_B

With 2 x_B bins



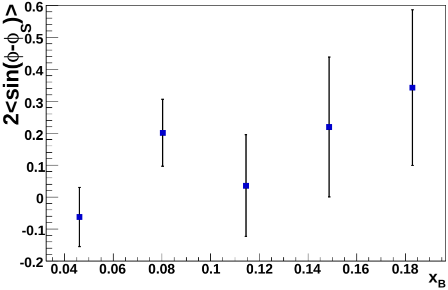
Fitting Monte Carlo data directly



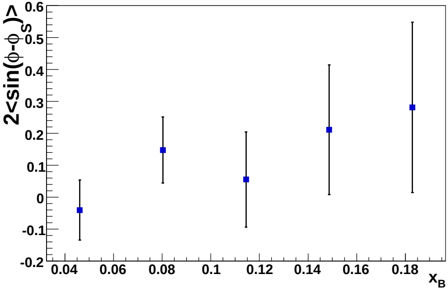
Fitting the 3D KDE, $h_x = 0.02$

- Denote bandwidth in x_B direction as h_x.
- Other bandwidths are $h_\phi = 2$, $h_{\phi_S} = 0.5$.
- Note: bandwidths not fully optimized, due to factors external to this example.
- With two x_B bins, no difference with or without using a KDE
- Cannot determine if dependence is statistically significant

With 5 x_B bins



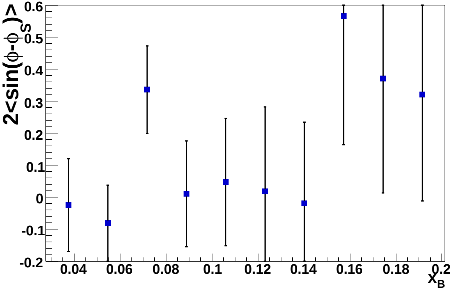
Fitting the data directly



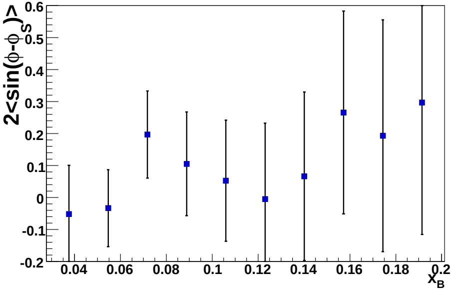
Fitting the 3D KDE, $h_x = 0.02$

- KDEs slightly "smoother"
- Note: KDEs are not considered "smoothing methods"
- KDEs accurately represent the data
- Full range of bandwidths yield KDEs from linear to delta functions

With 10 x_B bins



Fitting the data directly



Fitting the 3D KDE, $h_x = 0.02$

- Smoothness of KDE depends on bandwidth
- KDEs cannot overcome all difficulties of limited statistics
- This simple study does not include L/T separation, other details associated in actual analysis
- KDEs are additional tool for statistic samples—can be useful for other rare meson studies

Example 2: Unfolding

The Fredholm Integral Equation

- Measured distribution equals a smearing/acceptance operator acting on true distribution

$$p_{\mathcal{DV}}(\boldsymbol{x}^R) \;=\; \epsilon\kappa(\boldsymbol{x}^R) \int d\boldsymbol{x}^G \, p(\boldsymbol{x}^R|\boldsymbol{x}^G) p_{\mathcal{T}}(\boldsymbol{x}^G) \tag{7}$$

- PDF of measured data: $p_{\mathcal{DV}}(\boldsymbol{x}^R)$
- Smearing kernel is ratio of joint distribution to Born distribution, estimated using Monte Carlo data

$$p(\boldsymbol{x}^R|\boldsymbol{x}^G) = \frac{p_{\mathcal{MC}}(\boldsymbol{x}^R, \boldsymbol{x}^G)}{p_{\mathcal{MC}}(\boldsymbol{x}^G)}. \tag{8}$$

- ϵ is defined such that the right hand side integrates to 1.
- $\kappa(\boldsymbol{x}^R)$ accounts for any detector efficiencies not modeled by the Monte Carlo (often negligible)
- Unfolding is solving Equation 7 for the true distribution function $p_{\mathcal{T}}(\boldsymbol{x}^G)$, given data drawn from the densities $p_{\mathcal{DV}}(\boldsymbol{x}^R)$, $p_{\mathcal{MC}}(\boldsymbol{x}^R, \boldsymbol{x}^G)$, and $p_{\mathcal{MC}}(\boldsymbol{x}^G)$.
- Most numeric methods reduce integral equation to matrix equation $y = Ax$.

"Smeared-in Background"

- Note: $\mathcal{D}^R$, the domain of $\boldsymbol{x}^R$, is larger than $\mathcal{D}^G$, the $\boldsymbol{x}^G$ integration domain
- Separate true PDF into convex combination of PDFs over disjoint domains $\mathcal{D}_R$, $\mathcal{D}_G \backslash \mathcal{D}_R$.

$$p_{\mathcal{DV}}(\boldsymbol{x}^R) = \epsilon\kappa(\boldsymbol{x}^R) \int_{\mathcal{D}^G} d\boldsymbol{x}^G \, p(\boldsymbol{x}^R|\boldsymbol{x}^G) \begin{cases} \eta p_{\mathcal{T}}(\boldsymbol{x}^G) & \boldsymbol{x}^G \in \mathcal{D}^R \\ (1-\eta) p_{\mathrm{BKG}}(\boldsymbol{x}^G) & \text{otherwise} \end{cases} \tag{9}$$

- Rearrange to solve

$$p_{\mathcal{DV}}(\boldsymbol{x}^R) - \Upsilon(\boldsymbol{x}^R) p_{\mathrm{BKG}}(\boldsymbol{x}^R) = \kappa(\boldsymbol{x}^R)\epsilon\eta \int d\boldsymbol{x}^G \, p(\boldsymbol{x}^R|\boldsymbol{x}^G) p_{\mathcal{T}}(\boldsymbol{x}^G) \tag{10}$$

- Normalization $\Upsilon(\boldsymbol{x}^R)$ is defined to include all needed factors

Solving the Fredholm Equation

- Change

$$p_{\mathcal{DV}}(\boldsymbol{x}^R) = \epsilon \int d\boldsymbol{x}^G \, \frac{p_{\mathcal{MC}}(\boldsymbol{x}^R, \boldsymbol{x}^G)}{p_{\mathcal{MC}}(\boldsymbol{x}^G)} p_{\mathcal{T}}(\boldsymbol{x}^G) \quad \rightarrow \quad p_{\mathcal{DV}}(\boldsymbol{x}^R) = \epsilon \int d\boldsymbol{x}^G \, p_{\mathcal{MC}}(\boldsymbol{x}^R, \boldsymbol{x}^G) \frac{p_{\mathcal{T}}(\boldsymbol{x}^G)}{p_{\mathcal{MC}}(\boldsymbol{x}^G)} \tag{11}$$

- Use two basis expansions

$$R(\boldsymbol{x}^G) = \frac{\epsilon p_{\mathcal{T}}(\boldsymbol{x}^G)}{p_{\mathcal{MC}}(\boldsymbol{x}^G)} = \sum_k \zeta_k g_k(\boldsymbol{x}^G), \tag{12}$$

$$p_{\mathcal{T}}(\boldsymbol{x}^G) = \sum_i \alpha_i f_i(\boldsymbol{x}^G). \tag{13}$$

- Let $\boldsymbol{\beta} = \epsilon \boldsymbol{\alpha}$.
- The Fredholm equation can then be rewritten as

$$p_{\mathcal{DV}}(\boldsymbol{x}^R) = \int d\boldsymbol{x}^G \, p(\boldsymbol{x}^R, \boldsymbol{x}^G) \sum_k \zeta_k g_k(\boldsymbol{x}^G), \tag{14}$$

$$\sum_i \beta_i f_i(\boldsymbol{x}^G) = p_{\mathcal{MC}}(\boldsymbol{x}^G) \sum_k \beta_k g_k(\boldsymbol{x}^G). \tag{15}$$

Analytic Solutions

- Define:

$$A_{i,j} = \int d\boldsymbol{x}^G d\boldsymbol{x}^R \; e_i(\boldsymbol{x}^R) p_{\mathcal{MC}}(\boldsymbol{x}^R, \boldsymbol{x}^G) g_j(\boldsymbol{x}^G), \quad (16)$$

$$b_i = \int d\boldsymbol{x}^R \; e_i(\boldsymbol{x}^G) p_{\mathcal{DV}}(\boldsymbol{x}^G), \quad (17)$$

$$B_{i,j} = \int d\boldsymbol{x}^G f_i(\boldsymbol{x}^G) p_{\mathcal{MC}}(\boldsymbol{x}^G) g_j(\boldsymbol{x}^G), \quad (18)$$

$$D_{i,j} = \int d\boldsymbol{x}^G f_i(\boldsymbol{x}^G) f_j(\boldsymbol{x}^G), \quad (19)$$

$$c_i = \int d\boldsymbol{x}^G f_i(\boldsymbol{x}^G). \quad (20)$$

- Multiplying Equation 14 & 15 with $e_k(\boldsymbol{x}^R)$ and integrating over $\boldsymbol{x}^R$ yields

$$\boldsymbol{b} = A\boldsymbol{\zeta}, \qquad \boldsymbol{D\beta} = B\boldsymbol{\zeta}. \quad (21)$$

- Assuming A, D sufficiently well-conditioned and invertible, formal solution is

$$\widehat{\boldsymbol{\beta}} = D^{-1}BA^{-1}\boldsymbol{b}. \quad (22)$$

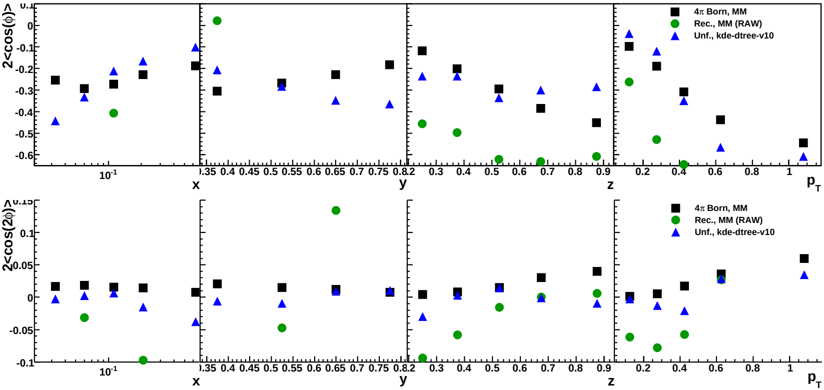
- Lastly, one can compute $\epsilon = \boldsymbol{c}^T\widehat{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{\alpha}} = \epsilon^{-1}\widehat{\boldsymbol{\beta}}$.
- Compute D, $\boldsymbol{c}$ analytically; estimate $A, \boldsymbol{b}, B$ by Monte Carlo Integration, summing over data drawn from respective density (generated from KDEs)
- Uncertainties can be propagated analytically

5D Monte Carlo Test

Density	Stats.
$p_{\mathcal{MC}}(\boldsymbol{x}^R, \boldsymbol{x}^G)$	4.5M
$p_{\mathcal{MC}}(\boldsymbol{x}^G)$	6.0M
$p_{\mathcal{DV}}(\boldsymbol{x}^R)$	1.2M
$p_{\mathcal{T}}(\boldsymbol{x}^G)$	11.6M

- Use LEPTO Monte Carlo to act as actual device (HERMES).
- Use PYTHIA Monte Carlo to act as Monte Carlo (as is done in analysis of real data).
- Basis set f_i chosen to be same 1D projections used for HERMES preliminary h^+/h^- $\cos(n\phi)$ moments.
- All basis sets $e_i = g_i$ are Cartesian product of $\cos(n\phi)$ moments (n=0,1,2) and piecewise constants, according to decision tree structure.
- Unfolding time on the order of 20 minutes (not including bandwidth optimization).
- Note: poor choices for kinematic portion of basis include those.
  - Too computationally expensive
    - KDEs, Splines
    - Multiple layers of kernels chosen to tessellate the domain
  - Inaccurate results
    - Piecewise affine (hyper-plane+const.)
    - Histograms (w/o “bad bin” removal)

Monte Carlo Results



- Decision tree in order p_T, z, y, x, dividing statistics of $p_{\mathcal{DV}}$ into 3rds at each level.
- KDEs used for $p_{\mathcal{DV}}$, $p_{\mathcal{MC}}(\boldsymbol{x}^G, \boldsymbol{x}^R)$, but not yet $p_{\mathcal{MC}}(\boldsymbol{x}^G)$.
- Systematic uncertainty still much larger than statistical—hope to improve with inclusion of $p_{\mathcal{MC}}(\boldsymbol{x}^G)$ KDE & further bandwidth optimization.
- Smeared-in background correction has been applied.
- Many other options for f_i—options for e_i, g_i limited by conditioned-ness of A.
- Can also extract kinematic properties, e.g. $\langle P_T \rangle$.

Conclusion

Conclusion

- KDE tools optimized for physics analysis developed
  - Although previous tools existed, extensive code developed/optimized for precision/accuracy in high D and w/ large statistics
  - Includes boundary conditions
  - Novel bandwidth optimization procedure
  - Evaluating KDEs and optimizing bandwidths relatively computationally intensive
  - Generating data from KDE very fast
  - All KDE code can be made publicly available, depending on the interest
- Points of Caution
  - May need to correct Fourier moments based on bandwidth
  - High dimensional functionals of non-parametric estimators often not feasible (must resort to basis functions)
  - Basis functions not needed for few dimensions nor more “simple” functionals

Conclusion

- Have shown KDEs w/ Basis Functions for
  - Azimuthal Moment Extraction with Small Statistics
  - 5D $\cos(n\phi)$ Unfolding
- KDEs also very promising for
  - Yet higher dimensional unfolding (6D for SIDIS A_{UT} moments)
  - Azimuthal Moment Extraction with Larger Statistics
  - Process Identification (SIDIS ρ^0 A_{UT})
  - Particle Identification
  - Monte Carlo Generation
  - ...
- Methods of solving integral inversion problems are applicable to other integral equations.
- Expect to see more KDEs in the future