

Kernels, Mixtures, Distances, Model Selection

Bruce G. Lindsay

June 4, 2002

Presented at Cleveland Workshop on
Developments and Challenges in Mixture
Models, Bump Hunting and Measurement
Error Models

1 Introduction

1.1 Project Objectives:

NONPARAMETRIC MODEL EVALUATION AND SELECTION using STATISTICAL DISTANCES

Data types: Develop methods for:
Continuous or **binary** sequences or **spherical** data.

Data : High dimensional $\mathbf{X} = (x_1, \dots, x_d)^0$, replicated n times $\mathbf{X}_1, \dots, \mathbf{X}_n$.

² **Large** d , even $d \sim n$.

Model Types Descriptive mixture model methods with emphasis on:

² Flexibility, weak assumptions, **but** structural relationships, clustering

Model Evaluation

- ² Model selection based on **Quality-of-fit** assessment (under vs. over) using distances

Key non-statistical goal:

- ² Ease and accuracy of **computation**. EG:
No d_j dimensional numerical integration!!

—

1.2 The Key ingredients

- ² **Kernel densities $K(x,y)$**
 - Build/Find special kernels suitable for easy calculations as densities and distances
- ² **Mixture sieves using special kernels**
 - Rich system in which all distributions can be approximated

² Quadratic distances using special kernels

– Numerically simple analogues to chi-squared distances

‣ But creates cells "automatically"

‣ #cells and locations replaced by tuning parameter h

² Model evaluation & selection strategies

– Using quadratic distances

– Today, minimum risk selection

² Collaborators: M. Markatou, S-C Chen, J. Liu, S. Ray, K. Yang

2 The Special Kernels

Example to keep in mind:

Example #1: The normal kernel:

$$K_{\eta}(\mathbf{x}, \mathbf{y}) = \frac{1}{(2\pi\eta)^{d/2}} e^{-\frac{1}{2\eta}(\mathbf{x} - \mathbf{y})^T(\mathbf{x} - \mathbf{y})}$$

Notice $\eta = \sigma^2$, the concentration parameter.

—

Topic Summary.

1. Describe the mathematical **assumptions** on the kernels that we use to construct the theory.
2. Describe Example #2, the **mutation kernel** for binary sequence data.
3. Describe Example #3, the **Poisson kernel** for spherical data.

—

Notation. Kernels $K(\mathbf{x}, \mathbf{y})$ will be integrated $d\lambda(\mathbf{x})$, where measure λ represents: Lebesgue on R^d , counting on $\{0,1\}^d$, uniform on S^d , sphere.

2.1 Basic Kernel assumptions.

Useful Heuristic Identify **kernel $K(\mathbf{x}, \mathbf{y})$** properties with **matrix \mathbf{K}** properties where $K(i, j) = K_{ij}$. Eg, integrations and matrix products:

$$\mathbf{K}g = \int K(\mathbf{x}, \mathbf{y}) g(\mathbf{y}) d\lambda(\mathbf{y})$$

2.1.1 Assumptions

The kernels of interest, $K(\mathbf{x}, \mathbf{y}) = K_{\eta}(\mathbf{x}, \mathbf{y})$, will satisfy

1. **symmetry** $K(\mathbf{x}, \mathbf{y}) = K(\mathbf{y}, \mathbf{x})$

2. **Positive definite:** $g^0 \mathbf{K} g \geq 0$

$$\int g(\mathbf{x}) K(\mathbf{x}, \mathbf{y}) g(\mathbf{y}) d\lambda(\mathbf{x}) d\lambda(\mathbf{y}) \geq 0$$

3. **Standardized:** $K(\mathbf{x}, \boldsymbol{\eta}) = \text{pdf in } x \text{ for } \boldsymbol{\eta}$
 defined: \mathbf{Z}

$$\int_{\mathbf{Z}} K(\mathbf{x}, \boldsymbol{\eta}) d\lambda(\mathbf{x}) = 1$$

- a. Call this the *kernel density* with vector parameter $\boldsymbol{\eta}$
- b. Call the prob measure $P_{\boldsymbol{\eta}}$.
- c. Note: Parameter space and sample space the same.
- d. Holds in example #1

4. **Concentration parameter.** There exists a *concentration parameter*

$$\eta \in (0, 1)$$

as a , so that

– as $\eta \rightarrow 0$

$\propto P_{(\boldsymbol{\eta}, \eta)}$ becomes **more** concentrated
 (IE, **point mass** δ_{μ})

– as $\eta \rightarrow 1$

$\propto P_{(\boldsymbol{\eta}, \eta)}$ becomes less concentrated (IE
 more "**uniform**")

\propto Obvious for example #1, $\eta = \sigma^2$.

2.1.2 Additive Kernels

Key assumption: **Product closure property (PCP)**. Say that the kernel has the **PCP** (or, is an **additive** kernel) in η if

$$K_{\eta_1} K_{\eta_2} = K_{\eta_1 + \eta_2}.$$

Iz, if:

$$K_{\eta_1}(\mathbf{x}, \mathbf{y}) K_{\eta_2}(\mathbf{y}, \mathbf{z}) d\lambda(\mathbf{y}) = K_{\eta_1 + \eta_2}(\mathbf{x}, \mathbf{z})$$

—

1. Ie, "matrix products" stay in same family
2. Concentration parameters add.
 - ² – (Note: sometimes requires reparameterization to construct an "**additive parameter**")

Why is additivity a good property?

Makes important integral calculations "easy".

—

2.1.3 Example #1: normal kernel

Normal kernel is **additive**. Reason: Here $K(\mathbf{x}, \mathbf{y}) = K(\mathbf{x} \mid \mathbf{y})$, so PCP comes from convolution formula:

$$\begin{aligned} X \mid \mu &\gg N(\mu, \sigma_1^2) \text{ and} \\ \mu &\gg N(z, \sigma_2^2) \\ \Rightarrow X &\gg N(z, \sigma_1^2 + \sigma_2^2) \end{aligned}$$

Additive parameter is $\eta = \sigma^2$.

3 Mixture models and sieves

Two ways to build a rich class of structured models using mixtures of additive kernels

—

Topic Summary:

1. Mixture models using $K_\eta(\mathbf{x}, \mathbf{1})$, written $K_\eta(\mathbf{x}, Q)$.
2. Two mixture sieves: η -sieve and k -sieve.
3. Properties of the sieves
4. NPMLE for mixture models (background)
5. η -sieve algorithm and resulting **mixture tree**
6. Example using genetic data

3.1 Mixture models

Let Q be a distribution on parameter space. Given *additive* kernel, construct **mixture density**

$$K_\eta(\mathbf{x}, Q) := \int K_\eta(\mathbf{x}, \mathbf{1}) dQ(\mathbf{1}).$$

(Note the device of using integrating measure as an argument.)

EG#1: Normal mean mixtures, with fixed σ^2 .

3.1.1 Two special cases:

1. **k-component mixture model:** Q discrete with k points of support, $Q_k = \sum_{j=1}^k \pi_j \delta_{\mathbf{1}_j}$

$$K_\eta(\mathbf{x}, Q_k) = \sum_{j=1}^k \pi_j K_\eta(\mathbf{x}, \mathbf{1}_j)$$

2. **Nonparametric model:** Make no assumptions about distribution/"parameter" Q .

3.1.2 Issues:

The full nonparametric mixture model

$$K_\eta(\mathbf{x}, Q)$$

² Is **not identifiable** over both η and Q together.

² Is **too rich** to fit using likelihood methods: get infinite density values in the limit, using \hat{F} for Q and $\eta = 0$.

² Hence we consider the following:

3.2 Two mixture sieves

3.2.1 k -Sieve:

For each k , maximize likelihood over (Q_k, η) .

As $k \uparrow$, get richer fits.

Stop when the fit is adequate, ie, before it fits too well.

3.2.2 η -Sieve:

For each η , maximize likelihood over Q nonparametrically.

As $\eta \downarrow 0$, get richer fits. Again, stop when fit is adequate.

3.2.3 Missing piece:

To do this, need to operationalize "**underfitting**" (poor fit because model is not rich enough) and "**overfitting**" (too good a fit, *a la* Fisher vs. Mendel).

3.2.4 Comparing the sieves

Parallels:

- ² For the k -sieve, k is fixed and $\hat{\eta}_k$ is random.
- ² For the η -sieve, η is fixed, but there exists a random, finite $\hat{k}_\eta =$ number of components in \hat{Q}_η .

—

Computation:

- ² Nonparametric MLE more complex to compute, but
 - More reliable due to uniqueness of solutions
 - Can be used to create a mixture tree

4 Quadratic distances

(with S. Ray, M. Markatou)

—

Topic Summary

1. Definition and description of quadratic distances
2. Role of the distance kernel. **Our choice** for K -mixtures is K_h , the same additive kernel
3. The **L2 distance** interpretation for additive kernels
4. **Selection** of the tuning parameter h using pDOF
5. **Estimating the distance** between truth and model.

4.1 Introduction

Definition. Given a *positive definite* kernel $\kappa(\mathbf{x}, \mathbf{y})$, define the **quadratic distance** $d(F, G)$ between two probability measures F and G to be: \mathbf{Z}

$$d(F, G) = \int \int \kappa(\mathbf{x}, \mathbf{y}) d(F; G)(\mathbf{x}) d(F; G)(\mathbf{y})$$

4.1.1 Interpretation:

Essentially a quadratic form in the **densities**, but defined w/o densities.

Calculation. Recall: Using integration notation, $\mathbf{Z} \mathbf{Z}$

$$\kappa(F, G) := \int \int \kappa(\mathbf{x}, \mathbf{y}) dF(\mathbf{x}) dG(\mathbf{y}).$$

Then need to calculate the following integrals:

$$d(F, G) = \kappa(F, F) + \kappa(F, G) + \kappa(G, F) + \kappa(G, G)$$

4.1.2 Properties:

- ² $d(F, G)$ is squared metric on p.m.s
- ² In discrete sample spaces, = quadratic form
- ² Mathy stuff: use Karhunen-Loeve decompositions. (Like eigenvector- eigenvalue analysis only for kernel functions)

—

4.1.3 Estimation:

Important: If F is discrete and G is continuous, some density-based distances (Kullback-Leibler) give infinite distance. **Not so** here, therefore

- ² In applications, can use $d(\hat{F}, M)$ to estimate $d(\tau, M)$, where τ is truth and M is a chosen model element.
- ² Quadratic nature makes it easy to **de-bias**.
- ² These estimated distances becomes tools in model selection

4.2 Choosing the distance kernel

4.2.1 Our choice

For the NP mixture model $K_\eta(\mathbf{x}, Q)$, use **same kernel** for distance:

$$\kappa = K_h(\mathbf{x}, \mathbf{y})$$

² Call h the *tuning* parameter for the kernel

² Additive kernel) the calculation of $d(\hat{F}, M)$ is "easy" for mixture model $M = K_\eta(\mathbf{x}, Q)$, provided Q is *discrete*. IE:

$$K_h(\mathbf{x}, M) = K_{h+\eta}(\mathbf{x}, Q) = \sum_j \pi_j K_{\eta+h}(\mathbf{x}, \mathbf{1}_j)$$

4.2.2 Issues in choice of κ :

² #1 is **ease of calculation**

– To calculate $d(F, G)$, avoid need for d -dimensional **numerical integration** or summation

– #2 is **sensitivity** of distance

– Use tuning parameter h to adapt

4.3 Selection of h

Problem: to select reasonable values for tuning parameter h .

4.3.1 Issues.

h selection problem **seems** similar to tuning parameter in **kernel density estimation**

BUT: Some important differences:

- ² Latter: $MSE = \text{Bias}^2 + \text{Variance}$. But here **bias** is not an issue.
 - ² Rather, seeking to discriminate between models. Smaller h , distance estimation more **noisy**, but **more sensitive** to local aberrations.
 - ² Seek a measure that indicates where we are on the Noise/sensitivity scale.
-

4.3.2 Pseudo degrees of freedom

Definition:

$$pDOF = \frac{\int_{\mathbf{R}} \int_{\mathbf{R}} K_h(\mathbf{x}, \mathbf{x}) d\tau(\mathbf{x})}{\int_{\mathbf{R}} \int_{\mathbf{R}} K_h^2(\mathbf{x}, \mathbf{y}) d\tau(\mathbf{x})d\tau(\mathbf{y})}$$

Interpretation

- ² If the estimated distance is asymptotically constant times chi-squared, gives **exact** degrees of freedom.
- ² Otherwise corresponds to **Satterthwaite's** two-moment based chi squared **approximation** to lim. distn.
- ² **Reason:** In terms of eigenvalues of K_h ,

$$pDOF = \frac{\sum \lambda_i}{\sum \lambda_i^2}$$

- ² Fits analogy of "**effective number of cells**" as follows: In $N(0, \sigma^2 I)$, for small h ,

$$pDOF \approx \frac{4\sigma^d}{2h}$$

This would be number of cells if the effective sample space was 4σ wide and the cells were $2h$ wide.

Estimation of pDOF

- ² Replace τ with \hat{F} .
- ² Use unbiased versions
- ² Initially, looks very promising

—

Future applications Develop some "rules of thumb". EG

- ² pDOF at least **3** to look for gross differences
- ² pDOF no more than **n/5** to retain adequate "averaging"
- ² Consider using **several values** depending on observed sensitivity to model selection.

5 Model selection problems

(with S. Ray, M Markatou, J. Liu)

Topic Summary

1. **Issues** in selecting models
2. Method 1: **density concordance** (informal, like R^2)
3. Method 2: **risk estimation**, using the quadratic distance as a loss function
4. Method 3: **distance estimation**, using normalized distance estimates to assess significant differences.
5. Method 4: Construction of **model "tubes"** to assess hypotheses about true distance.
6. To be developed: auxiliary **diagnostics** for model evaluation

5.1 The issues

Goal: **selection** of η or selection of k using statistical criterion.

5.1.1 Informal criteria:

Want methods that describe the η of the data, find clear structures, but recognize overfitting as "nothing left to η ". EG, R^2 in regression. (Note, like regression, quadratic distances have "units")

5.1.2 Formal criteria:

Target methods that

- η Give **lower confidence limits** for η or k parameters. (No upper limits possible)
- η are **consistent** for η or k
- η are **consistent for minimal adequate** η or k (allow for a certain amount of kernel failure)
- η give estimator with **minimal risk**

5.2 Method 2: Estimation of Risk

5.2.1 Strategy:

² Have a **method of estimation** of M , say \hat{M} , within each model (each η or k , here k)

² Create a nonnegative **global loss function**

$$L(\tau, \hat{M}_k)$$

that measures error in using \hat{M}_k in place of the **true distribution** τ . We will use

$$L(\tau, \hat{M}_k) = d_h(\tau, \hat{M}_k)$$

² Let the **risk** in using model k , when the true distribution is τ , be

$$R(k) = E_{\tau}[L(\tau, \hat{M}_k)]$$

² **Parallel:** Choosing a model to minimize risk is like choosing a model to minimize MSE for prediction, where $\text{MSE} = \text{Bias}^2 + \text{Variance}$.

² Corresponding **two** components to risk:

- i. If truth τ not in model, can never make loss zero. (**lack of fit**, like bias^2 , a cost due to underfitting)
- ii. And variability in \hat{M}_k (**parameter estimation error**, a cost due to overfitting)

² Tradeoff in risk as k increases:

- \propto Lack of fit decreases in k
- \propto Variability increases in k

² Construct a **risk estimator** $\hat{R}(k)$.

² **Pick** the model \hat{M}_k with the lowest estimated risk.

- **Goal:** select k in which the fitted model is likely to be closest to the true τ .)

5.2.2 Quadratic distance as loss

Fact, we can construct an **unbiased estimator** of $R_{n-2}(k)$ via

$$\hat{R}(k) = \frac{1}{n(n-1)} \sum_{i \neq j} K_{\langle i,j \rangle}(\mathbf{x}_i, \mathbf{x}_j)$$

where $M_{\langle i,j \rangle}$ is the "delete 2" model estimator and

$$K_{\langle i,j \rangle}(\mathbf{x}, \mathbf{y}) = K(\mathbf{x}, \mathbf{y}) + K(\mathbf{x}, M_{\langle i,j \rangle}) + K(M_{\langle i,j \rangle}, \mathbf{y}) + K(M_{\langle i,j \rangle}, M_{\langle i,j \rangle})$$

Note: as before, the additive property makes these calculations explicit.

In practice, we have been deleting more observations because the EM moves slowly.

5.2.3 Interpretation

Similar to *AIC* and *BIC*, but

- ² Parameter estimation costs not determined by simple **penalty** (here inoperative)
- ² Lack of Δ of model to data measured **directly**
 - Information criteria compare neighboring model Δ s, but use no **absolute** measure of Δ
- ² Speculation: More robust to discretization, other realities?

5.2.4 Surajit plots