Quantifying Uncertainty

Sai Ravela

M. I. T

Last Update: Spring 2013

Contents

Quick Recap

1. To model uncertainties in data, we represent it by probability density/mass.
2. These densities can be parametric forms, the exponential family is useful.
3. The parameters of the density functions may be inferred using a Bayesian approach.
4. It is particularly useful to use conjugate priors in the exponential family for the estimation of the density functionï¿½s parameters.

Density Estimation

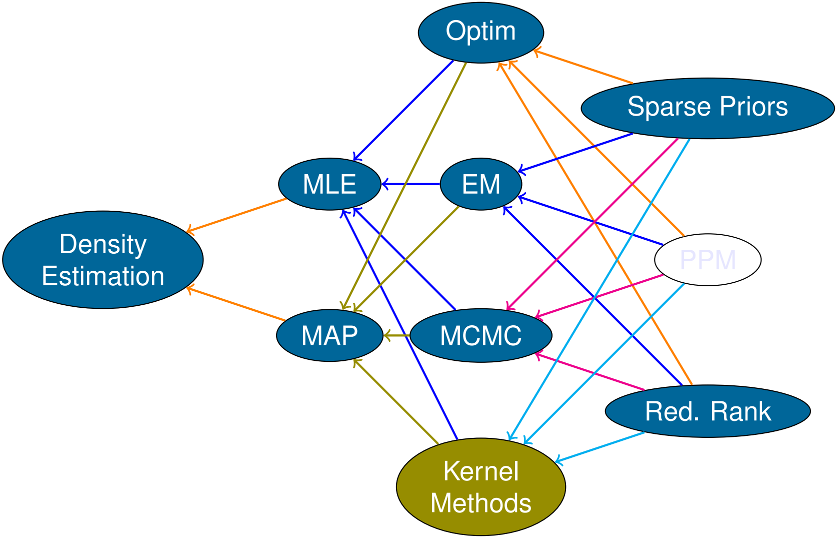
We want to estimate the parameters that control a Probability mass function P(Y = x;θ) from data.

For example, they could be the natural parameter in the exponential family of distributions, the mixing ratios in a mixture model etc.

A Bayesian approach to this problem would be to represent the unknown parameter as a random variable and consider its distribution i.e.

$$P(\theta|Y) \propto P(Y|\theta)P(\theta)$$

Methodological Space



Recall

Likelihood from Exponential Families:

$$l(\theta) = \sum \ln p(x_i|\theta) = \sum \ln h(x_i) + \theta^T T(x_i) - A(\theta)$$

$$\frac{dl}{d\theta} = 0 \Rightarrow \frac{1}{N} \sum_i T(x_i) = \frac{dA}{d\theta}$$

Example

$$p(x_i|\mu,\sigma) = 1/\sqrt{2\pi\sigma}e^{(x_i-\mu)^2/2\sigma^2}$$

$$\underline{\theta} = \begin{bmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{bmatrix}; T(x_i) = \begin{bmatrix} x_i \\ x_i^2 \end{bmatrix}; A(\underline{\theta}) = -\frac{\theta_1^2}{4\theta_2} - \frac{1}{2}\log(-2\theta_2)$$

$$dA/d\theta_1 = \frac{1}{N}\sum_i T_1 = \frac{1}{N}\sum_i x_i (= \mu)$$

$$dA/d\theta_2 = \frac{1}{N}\sum_i T_2 = \frac{1}{N}\sum_i x_i (= \mu^2 + \sigma^2)$$

Using an optimization

For more complicated distributions, some optimization procedure can be applied:
Let $F(\underline{\theta}) \doteq dA/d\underline{\theta}$ and $\frac{1}{N}\sum_i T(x_i) = \underline{z}$
Then solve: $||\underline{z} - F(\underline{\theta})||$
E.g. Levenberg-Marquardt:

$$J \doteq \frac{\partial F}{\partial \underline{\theta}}$$

Then,

$$[J^T J + \lambda tr(J^T J)]\delta\underline{\theta}^{(i)} = J^T(\underline{z} - F(\underline{\theta})^{(i)})$$

update, increment and iterate.
If you can easily calculate gradients, you could get fast (quadratic) convergence.

The Problem

Taking gradients is not always easy in closed form, and can be non-robust in numerical form especially with “noisy” likelihoods. What’s the alternative?
E.g. Mixture Density:

$$p(\underline{x}_i|\underline{\theta},\underline{\alpha}) = \sum_{s=1}^{s} \alpha_s G(\underline{x}_i;\underline{\theta}_s)$$

$$P(\chi|\underline{\theta},\underline{\alpha}) = \prod_{i=1}^{N} \sum_{s=1}^{s} \alpha_s G(\underline{x}_i;\underline{\theta}_s)$$

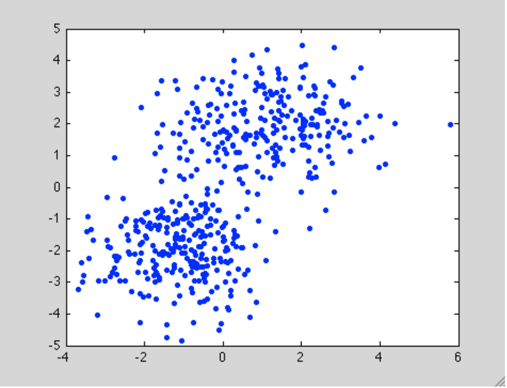
$$\begin{aligned} P(\underline{\theta},\underline{\alpha}|\chi) &\propto P(\chi|\underline{\theta},\underline{\alpha})P(\underline{\theta},\underline{\alpha}) \\ &= P(\underline{\theta},\underline{\alpha}) \prod_{i=1}^{N} \sum_{S=1}^{S} \alpha_s G(\underline{x}_i;\underline{\theta}_s) \end{aligned}$$

Contd.

$$J(\underline{\theta}, \underline{\alpha}) = \log P(\underline{\theta}, \underline{\alpha}|\chi) \ \propto \ \log P(\underline{\theta}, \underline{\alpha}) + \sum_{i=1}^{N} \log \left(\sum_{s=1}^{S} \alpha_s G(\underline{x}_i, ; \underline{\theta}_s) \right)$$

This is difficult, even when the prior is “trivial”

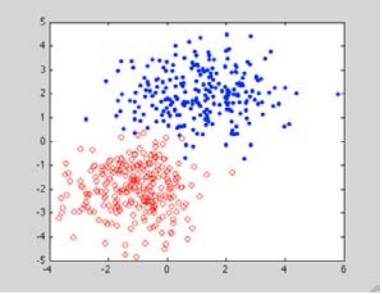
What's the mixture



Data given $x_i \in \chi$, what is its *pmf* (*pdf*)?
Mixture: How many members? Let's assume we know, even then:
What are the mixing proportions? Distribution parameters?

How can this problem be made easy?

What if someone tells you a key piece of missing information, i.e. which member of the distribution a data point comes from.



Then this is trivial!

What if?

We estimate an an expectation of the missing information, under a “complete distribution” that we also propose.

Then we maximize for the best set of parameters from that expectation.

We reuse the parameters to calculate a new expectation and keep iterating to convergence.

Thatï¿½s the EM algorithm in a nutshell.

Just what is the “complete distribution”, what are we expectating and what does all this converge to?

Formulation

$$p(\theta|X_i) \propto P(X_i|\theta)P(\theta)$$
$$\log P(\theta|X_i) \propto \underbrace{\log P(X_i|\theta)}_{1} + \log P(\theta)$$

$$1 \rightarrow \log P(X_i, \theta) = \log \sum_{yi} P(X_i, Y_i|\theta)$$

Introduce an iterative form:
Assuming an estimate of $\theta \Rightarrow \hat{\theta}^{(t)}$
So,

$$Q(\theta|\hat{\theta}^{(t)}) = \log P(\theta|X_i, \hat{\theta}^{(t)})$$

Contd.

Rewriting 1:

$$\log P(X_i|\theta) = \log \sum_{yi} P(X_i, Y_i|\theta)$$
$$= \log \sum_{yi} \frac{\phi(Y_i) P(X_i, Y_i|\theta)}{\phi(Y_i)}$$

$\phi(Y_i)$ is a distribution that “lower bounds” the likelihood

A. So, it must trade $P(X_i, Y_i|\theta)$ i.e

$$\frac{P(X_i, Y_i|\theta)}{\phi(Y_i)} = \kappa \text{ (some constant)}$$

Contd. 2

B. $\sum_{y_i} \phi(y_i) = 1$; it is a probability mass function

C. It exploits the availability of $\hat{\theta}^{(t)}$

$$\begin{aligned}\Rightarrow \phi(y_i) &= \frac{P(Y_i, X_i|\hat{\theta}^{(t)})}{\sum_y P(Y_i = y, X_i|\hat{\theta}^{(t)})} \\ &= \frac{P(Y_i, X_i|\hat{\theta}^{(t)}}{P(X_i|\hat{\theta}^{(t)})} \\ &= P(Y_i|X_i, \hat{\theta}^{(t)})\end{aligned}$$

So, it proposes a bound $\Rightarrow$ likelihood of missing data from most recent estimate of θ.

Contd. 3

$$Q(\theta|\hat{\theta}^{(t)}) = \log P(\theta) + \log \sum_{y_i} \frac{P(Y_i|X_i, \hat{\theta}^{(t)})P(X_i, Y_i|\theta)}{P(Y_i, |X_i, \hat{\theta}^{(t)})}$$

$$= \log P(\theta) + \log E\left[\frac{P(X_i, Y_i|\theta}{P(Y_i|X_i, \hat{\theta}^{(t)})}\right]$$

$$\geq \log P(\theta) + E \log\left[\frac{P(X_i, Y_i|\theta)}{P(Y_i|X_i, \hat{\theta}^{(t)})}\right]$$

$$H(Y_i|X_i, \hat{\theta}^{(t)})$$

$$= \log P(\theta) + E\left[\log P(X_i, Y_i|\theta)\right] - E\left[\log \cancel{p(Y_i, X_i, \hat{\theta}^{(t)})}\right]$$

$$\equiv \boxed{\log P(\theta) + E\left[\log P(X_i, Y_i|\theta)\right]}$$

Contd. 4

So, E-STEP:

$$\begin{aligned} Q(\theta|\hat{\theta}^{(t)})) &\equiv \log P(\theta) + E[\log P(X_i, Y_i|\theta)] \\ &= \log P(\theta) + \sum_{Y_i} P(Y|X_i, \hat{\theta}^{(t)}) \times \log P(X_i, Y_i|\theta) \end{aligned}$$

The prior + the expectation under $\hat{\theta}^{(t)}$ given, over missing variable Y_i.

Contd. 5

M-STEP

$$\hat{\theta}^{(t)} = ag \max_{\theta} Q(\theta|\hat{\theta}^{(t)})$$

alternate the two!
For many data samples $X_1, X_2, \ldots X_N \in \chi$

$$Q(\theta|\hat{\theta}^{(t)}) \equiv \sum_i \sum_{Y_i} P(Y_i|X_i, \hat{\theta}^{(t)}) \log P(X_i, Y_i|\theta) + \log P(\theta)$$

Measuring Similarity between Distributions

Kullback-Leibler Divergence

Distributions: $P(X)$, $Q(X)$

Divergence: $D(P\|Q) = \sum_x P(X) \log \frac{P(X)}{Q(X)}$

Interpretation: The cost of coding the “true” distribution $P(X)$ using a model distribution $Q(X)$

Interpretation:

$$
\begin{aligned}
D(P||Q) &= -\sum_x P(X) \log Q(X) - (-P(X) \log P(X)) \\
&= H(P, Q) - H(P)
\end{aligned}
$$

The relative entropy.

More

KL-Divergence is a broadly useful measure, for example:

Shannon Entropy: $H(X) = \log N - D(P(X)||U(X))$, departure from the uniform distribution.

Mutual Information: $(X; Y) = D(P(X, Y)||P(X)P(Y))$

Let's try to interpret EM in terms of KL divergence.

EM Interpretation

$$
\begin{aligned}
&\log P(\theta) + E \log \frac{P(X_i, Y_i|\theta)}{P(Y_i, X_i, \hat{\theta}^{(t)})} \\
&= \log P(\theta) - \sum P(Y_i|X_i, \hat{\theta}^{(t)}) \log \left[\frac{P(X_i, Y_i|\theta)}{P(Y_i|X_i, \hat{\theta}^{(t)})} \right] \\
&= \log P(\theta) + \log P(X_i|\theta) + \sum P(Y_i|X_i, \hat{\theta}^{(t)}) \log \left[\frac{P(Y_i|X_i, \theta)}{P(Y_i|X_i, \hat{\theta}^{(t)})} \right] \\
&\equiv \log P(\theta, X_i) - \sum P(Y_i|X_i, \hat{\theta}^{(t)}) \times \log \left[\frac{P(Y_i|X_i, \hat{\theta}^{(t)})}{P(Y_i|X_i, \theta)} \right]
\end{aligned}
$$

Contd.

$$\therefore Q(\theta|\hat{\theta}^{(t)}) = \log P(\theta|X_i) - \underbrace{\mathcal{D}(P(Y_i|X_i, \hat{\theta}^{(t)}) \| P(Y_i|X_i, \theta))}_{\substack{\text{KL-Divergence between estimates and} \\ \text{optimal conditional distributions} \\ \text{of missing data}}}$$

$$\mathcal{D} \to 0 \Rightarrow Q(\theta|\hat{\theta}^{(t)}) \to \log P(\theta|X_i) \qquad (\text{Recall, } \mathcal{D} \geq 0)$$

Notes

1 The M-step can produce any $\hat{\theta}^{t+1}$ that improves Q, not just the maximum (at each iteration). That's Generalized EM (GEM).

2 M can be simpler to formulate for an MLE problem, and easier to implement than gradient-based methods. A huge explosion of applications, as a result.

   In applications of mixture modeling, EM method is synonymous with density estimation.

3 Convergence can be slow, i.e. if you can do Newton-Raphson (for example), do it.

What does it converge to?

Recall: $\mathcal{D}(P\|Q) \geq 0, \quad \mathcal{D}(P\|P) = 0. \quad \mathcal{D}(Q\|Q) = 0$

$$Q(\theta|\hat{\theta}^{(t)} = \sum_i \log P(\theta|X_i) - \mathcal{D}[P(Y_i|X_i, \hat{\theta}^{(t)})\|P(Y_i|X_i, \theta)]$$

$$\therefore Q(\hat{\theta}^{(t)}|\hat{\theta}^{(t)}) = \sum_i \log P(\hat{\theta}^{(t)}|X_i)$$

$$Q(\hat{\theta}^{(t)}|\hat{\theta}^{(t)}) = \sum_i \log P(\hat{\theta}^{(t+1)}|X_i) - \mathcal{D}[P(Y_i|X_i, \hat{\theta}^{(t)})\|P(Y_i|X_i, \hat{\theta}^{(t+1)})]$$

Contd.

$$\underbrace{Q(\hat{\theta}^{(t+1)}|\hat{\theta}^{(t)})}_{Q_{t+1}} \geq \underbrace{Q(\hat{\theta}^{(t)}|\hat{\theta}^{(t)})}_{Q_t}, \text{ by construction}$$

$$Q_{t+1} - Q_t \geq 0$$

$$\sum_i \log P(\hat{\theta}^{(t+1)}|X_i) - \log P(\hat{\theta}^{(t)}|X_i) \geq \mathcal{D}[P(Y_i|X_i, \hat{\theta}^{(t)}) \| P(Y_i|X_i, \hat{\theta}^{(t+1)}]$$

$$\geq 0$$

Posterior improves !

Stationary Points

$$\frac{dQ(\theta|\hat{\theta}^{(t)})}{d\theta}|_{t=\infty} =$$

$$\sum_i \frac{\partial \log P(\theta|X_i)}{\partial\theta}|_{\theta=\hat{\theta}^{(t)}} - \cancelto{0}{\frac{\partial \mathcal{D}[P(Y_i|X_i,\hat{\theta}^{(\infty)})\|P(Y_i|X_i,\theta)]}{\partial\theta}|_{\theta=\hat{\theta}^{(\infty)}}}$$

$$\Rightarrow \sum_i \frac{\partial \log P(\theta|X_i)}{\partial\theta}|_{\theta=\hat{\theta}^{(\infty)}} = 0$$

A stationary point of a posterior.

Gaussian Mixture Model

$$\prod_{i=1}^{N}\left[\sum_{s=1}^{S}\alpha_s P(X_i|\theta_s)\right] \times P(\theta_s) \qquad //MAP$$

$$\alpha \equiv \sum_{i=1}^{N}\log\sum_{s=1}^{S}\alpha_s P(X_i|\theta_s) + \log P(\theta_s)$$

only MLE for now

$$\geq \sum_{i=1}^{N}\sum_{s=1}^{S}\log[\alpha_s P(X_i|\theta_s)]$$

How to solve?

Contd.

Suppose there is an indicator variable $Y_{i,s}$
$Y_{i,s} \in \{0, 1\}$ and it is 1 when data X_i is drawn from distribution θ_s, then "total" likelihood (including "missing" data $Y_{i,s}$)

$$\alpha_{TOT} \equiv \sum_{i=1}^{N} \sum_{s=1}^{S} Y_{i,s} \log[\alpha_s P(X_i|\theta_s)]$$

we have to add <u>constraint</u> $\sum_j \alpha_j = 1$, so

$$\mathcal{L}_{TOT} + \lambda \left[\sum_j \alpha_j - 1 \right] = L$$

Differentiating, we get:

$$\frac{\partial L}{\partial \alpha_s} = \sum_{i=1}^{N} \frac{Y_{i,s}}{\alpha_s} + \lambda = 0$$

$$\Rightarrow \hat{\alpha}_s = \frac{\sum_{i=1}^{N} Y_{i,s}}{-\lambda}$$

Because

$$\sum_{i=1}^{N} Y_{i,s} + \lambda \alpha_s = 0 \quad \forall s$$

$$\therefore \sum_{s=1}^{S} \sum_{i=1}^{N} Y_{i,s} + \lambda \alpha_s = 0$$

$$\sum_{s=1}^{S} \alpha_s = 1, \quad \sum_{s=1}^{S} Y_{i,s} = 1 \Rightarrow \therefore \begin{cases} -\lambda = N \\ \text{or } \boxed{\hat{\alpha}_s = \frac{\sum_{i=1}^{N} Y_{i,s}}{N}} \end{cases}$$

Contd.

And

$$\hat{\theta}_s \equiv arg \max_{\theta_s} \sum_{i=1}^{N} Y_{i,s} \log(\alpha_s P(X_i|\theta_s))$$

No interaction between mixture elements given $Y_{i,s}$!
But we do not know $Y_{i,s}$, we estimated it through

$$P(Y_{i,s}|X_i, \underbrace{\hat{\theta}_s^{(t)}, \hat{\alpha}_s^{(t)}}_{\substack{\textit{Current}\\ \textit{estimates}}})$$

Contd.

We need to define: $Q(\underline{\theta}, \alpha | \hat{\underline{\theta}}^{(t)}, \hat{\alpha}^{(t)})$

$$P(Y_{i,s} | X_i, \hat{\underline{\theta}}_s^{(t)}, \hat{\underline{\alpha}}_s^{(t)}) = \boxed{w_{i,s} = \frac{\hat{\underline{\alpha}}_s^{(t)} P(X_i | \hat{\underline{\theta}}_s^{(t)})}{\sum_r \hat{\underline{\alpha}}_r^{(t)} P(X_i | \hat{\underline{\theta}}_r^{(t)})}}$$

$$Q \equiv \sum_{i=1}^{N} \sum_{s=1}^{S} w_{i,s} \log \frac{\alpha_s P(X_i | \underline{\theta}_s)}{w_{i,s}} + \lambda \left(\sum_j \alpha_j - 1 \right) \quad \text{// } w_{i,s} \text{ lower bounds}$$

$$\frac{\partial Q}{\partial \alpha_s} = \sum_{i=1}^{N} \frac{w_{i,s}}{\alpha_s} + \lambda = 0$$

$$\Rightarrow \boxed{\hat{\alpha}_s^{(t+1)} = \frac{\sum_{i=1}^{N} w_{i,s}}{N}}$$

Contd.

From exponential Family:

$$\log P(X_i|\underline{\theta}_s) = \log h(x) + \underline{\theta}_S^T T(X_i) - A(\underline{\theta}_s) \text{ // exponential family}$$

$$\frac{dQ}{d\underline{\theta}} = \sum_{i=1}^{N} w_{i,s} T(X_i) - \sum_{j=1}^{N} w_{j,s} \frac{dA}{d\underline{\theta}_s}$$

$$\Rightarrow \frac{dA}{d\underline{\theta}_s} = \frac{\sum_{i=1}^{N} w_{i,s} T(X_i)}{\sum_{i=1}^{N} w_{i,s}}$$

Contd.

So,

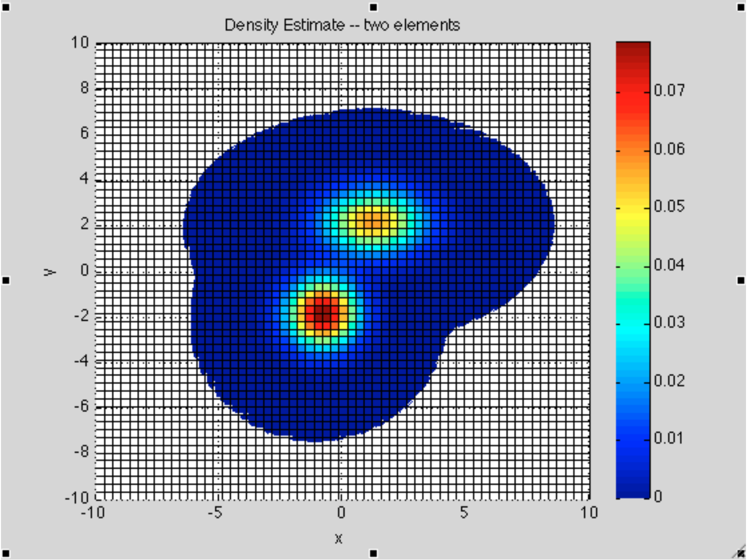
$$\frac{dA}{d\theta_1} = 0 \Rightarrow \underline{\hat{\mu}}^{t+1} = \frac{\sum_{i=1}^{N} w_{i,s} X_i}{\sum_{i=1}^{N} w_{i,s}}$$

$$\frac{dA}{d\theta_2} = 0 \Rightarrow \hat{\Sigma}^{t+1} + [\hat{\mu}\hat{\mu}^T]^{t+1} = \frac{\sum_{i=1}^{N} w_{i,s} X_i X_i^T}{\sum_{i=1}^{N} w_{i,s}}$$

Recall

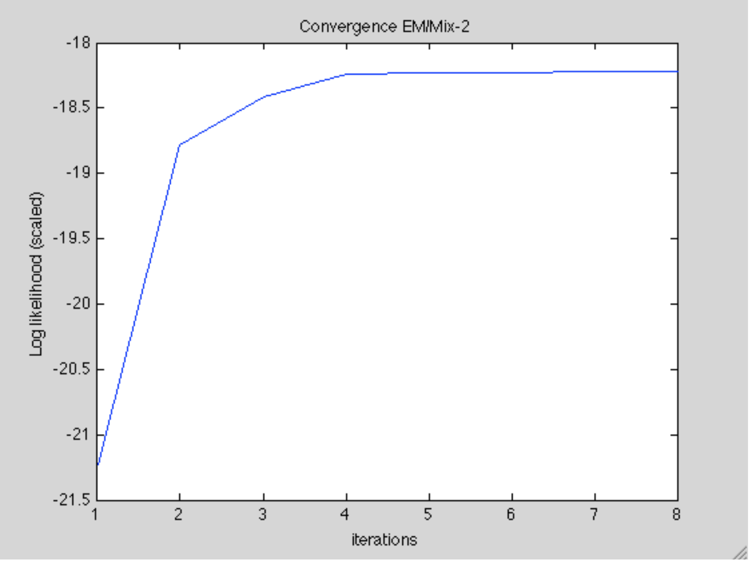
$$T(\alpha_1) = \begin{bmatrix} X_i \\ X_i X_i^T \end{bmatrix}$$

Example



Density Estimate -- two elements
10
8
6
4
2
0
-2
-4
-6
-8
-10
y
-10
-5
0
5
10
x
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0

Convergence



Model selection

How do we know how many members exist in the mixture? How to estimate it?

What is the best model to pick?

mpirical: Bootstrap, Jacknife, Cross-validation.

Algorithmic: AICc, BIC, MDL, MML (there are others, e.g. SRM).

Cross-Validation

You produce K sample sets, train on K-1, test on the remaining. Do this in turn. The simplest way to estimate parameter uncertainty, and produce somewhat robust result.

For 2-way cross-validation, you get the classical “Train & Test” data sets.

Algorithmic Approach

$P(\chi|\theta)$ is the true likelihood - of some "perfect" representation of the data.

$Q(\chi|\theta_n)$ is the approximate likelihood - of a model of the data. We want to figure out if Q is any good.

If we knew $P(\chi|\theta)$, we could calculate the KL-Divergence

$$D(P||Q) = \sum_x P(\chi = x|\theta) \log \frac{P(\chi = x|\theta)}{Q(\chi = x|\theta)} = H(P, Q) - H(P)$$

So, we may minimize "cross-entrop" or maximize ?

$$E_p[\log Q(\chi = x|\theta_n)]$$

Will this work?

Take 2

Let's assume smoothness in Q and take a Taylor Expansion:
$\log Q(\chi, \theta_n) \doteq L(\chi, \theta_n)$

$$\begin{aligned} L(\chi, \theta_n) &= L(\chi, \hat{\theta}_n) + (\theta_n - \hat{\theta}_n)^T \left. \frac{\partial L}{\partial \theta} \right|_{\theta = \theta_n^0} \\ &+ \frac{1}{2} (\theta_n - \hat{\theta}_n)^T \frac{\partial^2 L}{\partial \theta^2} (\theta_n - \hat{\theta}_n) \end{aligned}$$

An Information Criterion

$$
\begin{aligned}
E_p[L(\chi,\theta_n)] &= E_p[L(\chi,\hat{\theta}_n)] - E_p\left[\frac{1}{2}(\theta_n^0-\hat{\theta}_n)^T\sum^{-1}(\theta_n^0-\hat{\theta}_n)\right] \\
&= E_p[L(\chi,\hat{\theta}_n)] - E_p\left[\frac{1}{2}\sum^{-1}(\theta_n^0-\hat{\theta}_n)(\theta_n^0-\hat{\theta}_n)^T\right] \\
&= E_p[L(\chi,\hat{\theta}_n)] - Tr\left[\frac{1}{2}I_n\right] \\
&= E_p[L(\chi,\hat{\theta}_n)] - Tr\left[\frac{1}{2}n\right]
\end{aligned}
$$

An unbiased estimate is : $L(\chi,\hat{\theta}_n) - \frac{1}{2}n$

Giving a criterion: $-L(\chi,\hat{\theta}_n) + 2n$, for which we seek minimum.

For Gaussian: $N \ln \sigma^2 + 2n$ (N=number of samples, n=size of model, e.g. number of mixtures)

Akaike Information Criterion (AIC)

OK, but we don't know Ep; so, we cross-validate. Let's assume we have an independent data set from which we estimate parameters $\theta_n^{(x)}$

We write out the log-likelihood as $\ln P(\cdot) = E_x \ln Q(\chi, \theta_n^{(x)})$ and evaluate $E_p(\ln P) = E_p(E_x(\ln Q(\chi, \theta_n^{(x)})))$

This gives the AIC criterion: $-2L(\chi, \hat{\theta}_n) + 2n$

Others

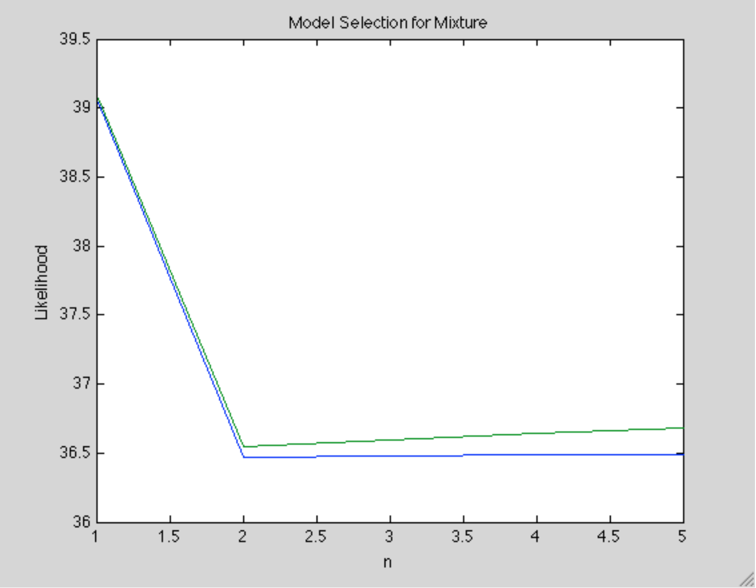
AICc: Correction to AIC for small samples: $-2L(\chi, \hat{\theta}_n) + \frac{2N}{N-n-1}n$

BIC: $-2L(\chi, \hat{\theta}_n) + n \ln N$

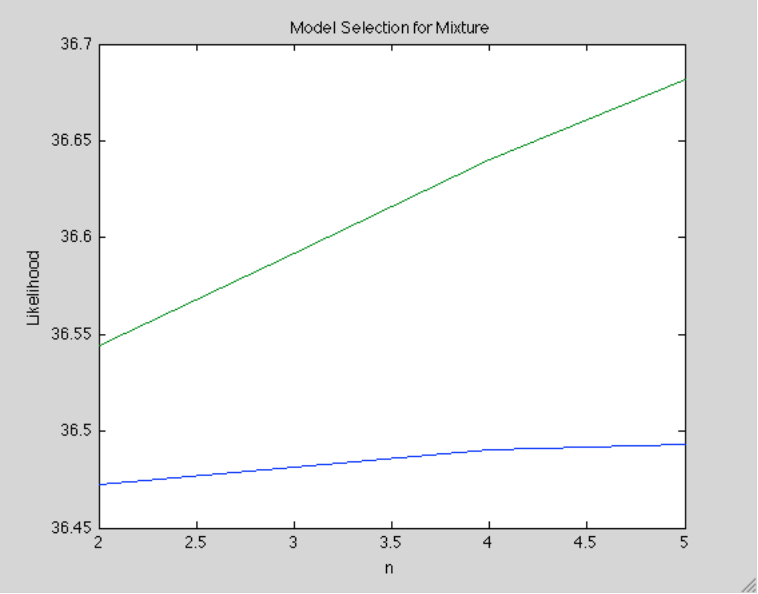
There are other information theoretic criteria, not covered here: MDL (Minimum Description Length) and MML (Minimum Message Length) are both powerful.

Model Selection is not a settled question! You should try multiple model selection criterion and evaluate.

Example


Model Selection for Mixture
Likelihood
n
39.5
39
38.5
38
37.5
37
36.5
36
1
1.5
2
2.5
3
3.5
4
4.5
5

Zoomed


Model Selection for Mixture
Likelihood
36.7
36.65
36.6
36.55
36.5
36.45
2
2.5
3
3.5
4
4.5
5
n

12.S990 Quantifying Uncertainty
Fall 2012