FRIEDRICH-ALEXANDER-
UNIVERSITÄT
ERLANGEN-NÜRNBERG

SCHOOL OF ENGINEERING

Lecture Pattern Analysis

Part 10: Model Selection for GMMs

Christian Riess
IT Security Infrastructures Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg
May 30, 2021

Introduction

- We use samples $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ and latent variables $\mathbf{Z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_N\}$
- Let us rewrite the GMM equations with explicit parameters,

$$p(\mathbf{Z}|\boldsymbol{\pi}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}} \tag{1}$$

and

$$p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\mu}, \Lambda) = \prod_{n=1}^{N} \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \Lambda_k^{-1})^{z_{nk}} \quad \text{where} \quad \Lambda = \Sigma^{-1} \tag{2}$$

- The Bayesian way of model selection is to add priors to the parameters
- Hence, $\boldsymbol{\pi}$, $\boldsymbol{\mu}$, Λ require suitable priors
- Exact inference is oftentimes not possible, popular approximations are Variational Inference[1] and Markov Chain Monte Carlo (MCMC)[2]

[1] If you like to know more: Variational Inference is covered in Bishop Sec. 10

[2] If you like to know more: MCMC is covered in Bishop Sec. 11

Remarks on Priors

- Priors for $\boldsymbol{\mu}_i$, Λ_i, π_i allow to "draw" new components, and to regularize fits on limited data
- Important: if likelihood and prior are **conjugate**, their product is from the same family of distributions as the likelihood
- Conjugate priors make the calculation **much** easier — always choose conjugate priors!.
- Priors bring additional parameters
  - The prior often has one more parameter than its associated likelihood — so nothing gained?
  - Not quite: the influence of the prior's parameters is very indirect, and further reduces with increasing dataset size
  - Hence, these parameters can be rather generic ("1", "mean of the data", ...)

Specific Priors: Dirichlet Distribution for the Mixing Coefficients

- The mixing coefficients $\boldsymbol{\pi}$ forms a **multinomial distribution** that can be seen as the relative number of samples that belong to a component
- The conjugate prior of the multinomial distribution is the Dirichlet distribution[3]

$$\text{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}_0) = C(\boldsymbol{\alpha}_0) \prod_{k=1}^{K} \pi_k^{\alpha_k - 1} \tag{3}$$

with k identical positive values α_0 in $\boldsymbol{\alpha}_0$, i.e., $\boldsymbol{\alpha}_0 = (\alpha_0, ..., \alpha_0) \in \mathbb{R}_+^k$, and a scaling factor

$$C(\boldsymbol{\alpha}_0) = \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\Gamma(\alpha_1) \cdot \ldots \cdot \Gamma(\alpha_k)} \tag{4}$$

[3] A list of distributions together with short descriptions and their role as conjugate prior (if applicable) is in Appendix B in the book by Bishop

Specific Priors: Gaussian-Wishart for Mean and Precision Matrix

- The prior for mean and precision matrix Λ (recall $\Lambda = \Sigma^{-1}$) is

$$p(\boldsymbol{\mu}, \Lambda) = p(\boldsymbol{\mu}|\Lambda)p(\Lambda) \tag{5}$$

$$= \prod_{k=1}^{K} \mathcal{N}\left(\boldsymbol{\mu}_k|\mathbf{m}_0, (\beta_0\Lambda_k)^{-1}\right) \mathcal{W}(\Lambda_k|\mathbf{W}_0, \nu_0) \tag{6}$$

where $\mathbf{m}_0$, β, $\mathbf{W}_0$, ν_0 are hyperpriors, and[4]

$$\mathcal{W}(\Lambda_k|\mathbf{W}_0, \nu_0) = B(\mathbf{W}_0, \nu_0)|\Lambda|^{(\nu_0 - D - 1)/2} \exp\left(-\frac{1}{2}\mathrm{Tr}(\mathbf{W}_0^{-1}\Lambda)\right) \tag{7}$$

with feature dimensionality D and the normalizing constant

$$B(\mathbf{W}_0, \nu_0) = |\mathbf{W}_0|^{-\nu_0/2} \left(2^{\nu_0 D/2} \pi^{D(D-1)/4} \prod_{i=1}^{D} \Gamma\left(\frac{\nu_0 + 1 - i}{2}\right)\right)^{-1} \tag{8}$$

[4] Formally, the Wishart distribution is the distribution of sample covariance matrices. As stated also later in this lecture, you do not need to memorize this equation.

Joint Distribution and Variational Approximation

- With our priors, the joint GMM distribution can be written as

$$p(\mathbf{X}, \mathbf{Z}, \pi, \mu, \Lambda) = p(\mathbf{X}|\mathbf{Z}, \pi, \mu, \Lambda)p(\mathbf{Z}|\pi)p(\pi)p(\mu|\Lambda)p(\Lambda) \tag{9}$$

- We use Bishop's variational framework (Sec. 10-10.2.1) for inference
- Note that this is just an illustration. The details are omitted on purpose[5].
- Bishop's variational framework approximates the distribution $p(\mathbf{X}, \mathbf{Z}, \pi, \mu, \Lambda)$ using a distribution $q(\mathbf{Z}, \pi, \mu, \Lambda)$ with the independence assumption

$$q(\mathbf{Z}, \pi, \mu, \Lambda) = q(\mathbf{Z})q(\pi, \mu, \Lambda) \tag{10}$$

- After several calculations, the solution is again an EM algorithm

[5] However, I cordially invite those of you who are curious to know more to a special meeting, where we can go through this Section in full detail

EM Solution: Expectation Step

- Calculate responsibilities

$$r_{ik} = \mathbb{E}[z_{nk}] = \frac{\rho_{nk}}{\sum_{j=1}^{K} \rho_{nj}} \tag{11}$$

where

$$\rho_{nk} = \mathbb{E}[\ln \pi_k] + \frac{1}{2}\mathbb{E}[\ln |\Lambda_k|] - \frac{D}{2}\ln(2\pi) - \frac{1}{2}\mathbb{E}_{\boldsymbol{\mu}_k,\Lambda_k}[(\mathbf{x}_n - \boldsymbol{\mu}_k)^\mathsf{T}\Lambda_k(\mathbf{x}_n - \boldsymbol{\mu}_k)] \tag{12}$$

- Our EM algorithm for standard GMM fitting looks somewhat simpler.
- However, the main difference here is that we operate on expectations over distributions (which are induced from the priors).
- The next slide lists the expanded equations for the expectations, but again only for illustration

EM Solution: Expectation Step / Expanded Equations

- Expanded equations for the expectations:

$$\mathbb{E}_{\boldsymbol{\mu}_k,\Lambda_k}[(\mathbf{x}_n-\boldsymbol{\mu}_k)^{\mathsf{T}}\Lambda_k(\mathbf{x}_n-\boldsymbol{\mu}_k)] = D\beta_k^{-1} + \nu_k(\mathbf{x}_n-\mathbf{m}_k)^{\mathsf{T}}\mathbf{W}_k(\mathbf{x}_n-\mathbf{m}_k) \tag{13}$$

$$\mathbb{E}[\ln|\Lambda_k|] = \sum_{i=1}^{D}\psi\left(\frac{\nu_k+1-i}{2}\right) + D\ln 2 + \ln|\mathbf{W}_k| \tag{14}$$

$$\mathbb{E}[\ln\pi_k] = \psi(\alpha_k) - \sum_{k=1}^{K}\alpha_k \tag{15}$$

where $\psi(\alpha)$ is the digamma function,

$$\psi(\alpha) = \frac{\mathrm{d}}{\mathrm{d}\alpha}\ln\Gamma(\alpha) \tag{16}$$

EM Solution: Maximization Step / Mixing Coefficients

- The distribution of mixing coefficients is updated via the Dirichlet distribution

$$q^*(\boldsymbol{\pi}) = \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}) \tag{17}$$

where each component α_k of $\boldsymbol{\alpha}$ is updated with the sum of its responsibilities

$$\alpha_k = \alpha_0 + N_k \tag{18}$$

with

$$N_k = \sum_{n=1}^{N} r_{nk} \tag{19}$$

EM Solution: Maximization Step / Means and Precision

- The distribution of means and precision matrices is updated with

$$q^*(\boldsymbol{\mu}_k, \Lambda_k) = \prod_{k=1}^{K} \mathcal{N}\left(\boldsymbol{\mu}_k | \mathbf{m}_k, (\beta_k \Lambda_k)^{-1}\right) \mathcal{W}(\Lambda_k | \mathbf{W}_k, \nu_k) \tag{20}$$

where

$$\beta_k = \beta_0 + N_k \tag{21}$$

$$\nu_k = \nu_0 + N_k + 1 \tag{22}$$

$$\mathbf{m}_k = \frac{1}{\beta_k}(\beta_0 \mathbf{m}_0 + N_k \bar{\mathbf{x}}_k) \tag{23}$$

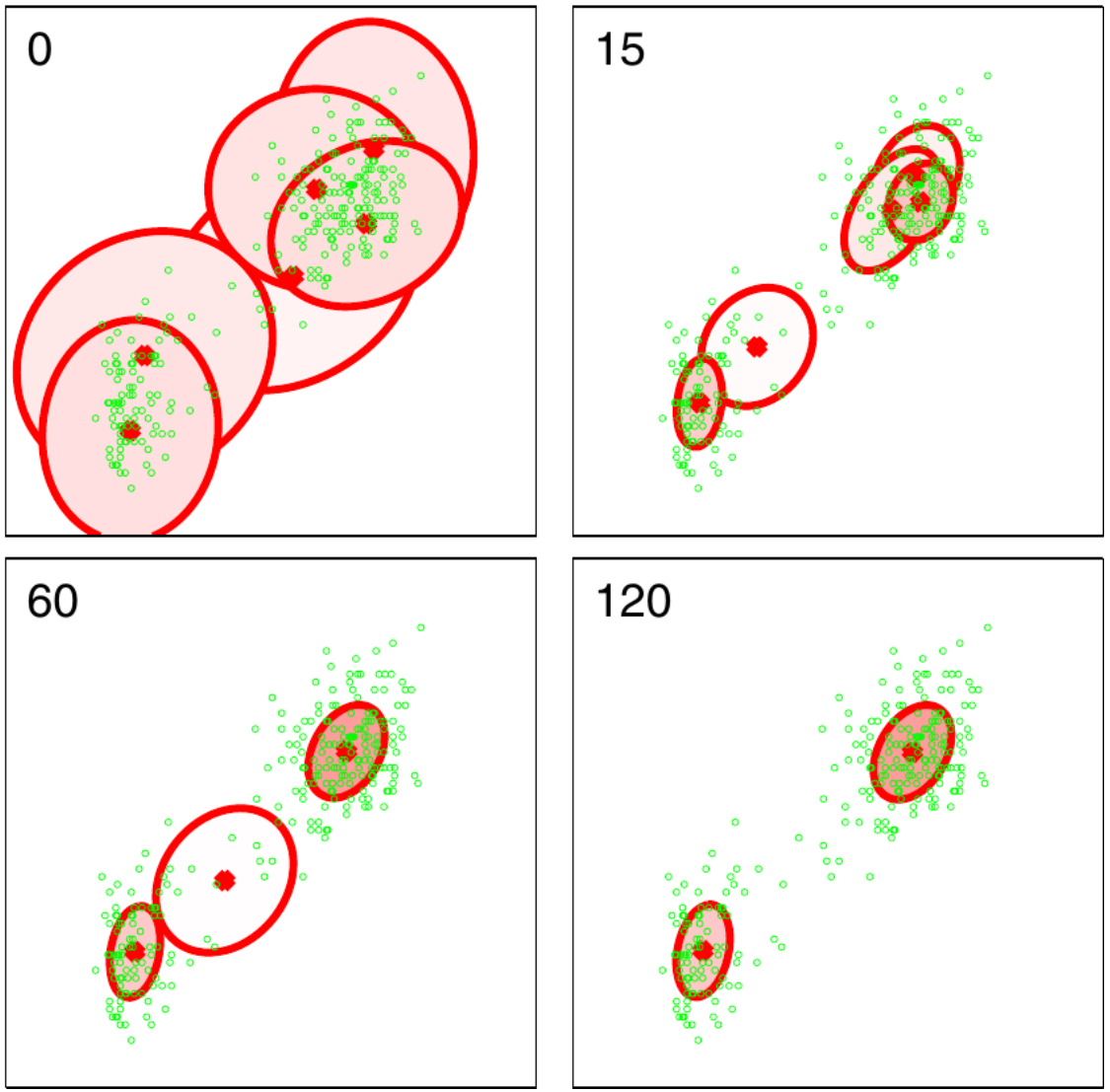
$$\bar{\mathbf{x}}_k = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} \mathbf{x}_n \tag{24}$$

$$\mathbf{W}_k^{-1} = \mathbf{W}_0^{-1} + N_k \mathbf{S}_k + \frac{\beta_0 N_k}{\beta_0 + N_k}(\bar{\mathbf{x}}_k - \mathbf{m}_0)(\bar{\mathbf{x}}_k - \mathbf{m}_0)^\mathsf{T} \tag{25}$$

$$\mathbf{S}_k = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk}(\mathbf{x}_n - \bar{\mathbf{x}}_k)(\mathbf{x}_n - \bar{\mathbf{x}}_k)^\mathsf{T} \tag{26}$$

Example Fit

Figure 10.6 Variational Bayesian mixture of $K = 6$ Gaussians applied to the Old Faithful data set, in which the ellipses denote the one standard-deviation density contours for each of the components, and the density of red ink inside each ellipse corresponds to the mean value of the mixing coefficient for each component. The number in the top left of each diagram shows the number of iterations of variational inference. Components whose expected mixing coefficient are numerically indistinguishable from zero are not plotted.



Remarks

- The priors can be thought of as regularizers that prevent overfitting
- General behavior:
  - Few data points: the priors dominate the result
  - Many data points: the data dominates the result
- Unnecessary components are automatically removed:
  α_k with (almost) zero responsibility approaches the uniform start value α_0
- Hence, it makes sense to start with a larger-than-expected number of clusters, and to remove at the end those clusters with $\alpha_k \approx \alpha_0$

- Markov Chain Monte-Carlo methods are an alternative for Bayesian inference
- One specific example is Gibbs sampling. Here,
  1. Randomly select one sample
  2. Calculate the conditional distributions of the parameters without that sample
  3. Randomly assign the sample to an existing or new cluster with probability of the cluster likelihood
  4. Goto 1. (theoretically until infinity)