Lecture Pattern Analysis

Part 09: Model Selection for K-Means

Christian Riess
IT Security Infrastructures Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg
19. Mai 2025

Introduction

- Clustering is unsupervised, and does not provide an objective function for model selection
- So, specifically for k-means: what k shall we choose?
- Even if the application demands, e.g., the “3 most important clusters”, $k = 3$ could be a poor choice if the intrinsic number of clusters is larger
- In this lecture, we investigate the **Gap Statistics** as a statistical way to determine k[1]
- The idea is to
  - examine the k-means optimization criterion, the **Within-Cluster Distance** $W(C)$, for different k,
  - and to select the smallest k for which $W(C)$ is substantially better than the $W(C)$ of $k + 1$ clusters

[1] The gap statistics is covered in the book by Hastie/Tibshirani/Friedman Sec. 14.3.11

Examining the Within-Cluster-Distance $W(C)$

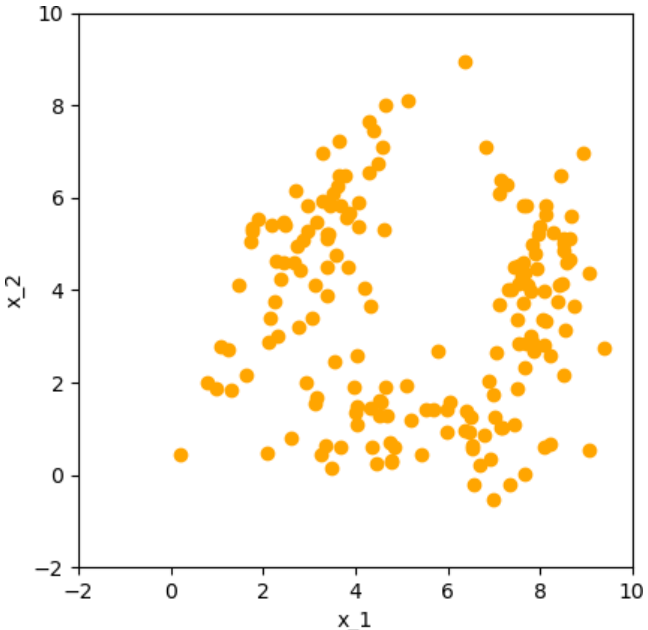
- Recall that we defined the Within-Cluster Distance $W(C)$ as

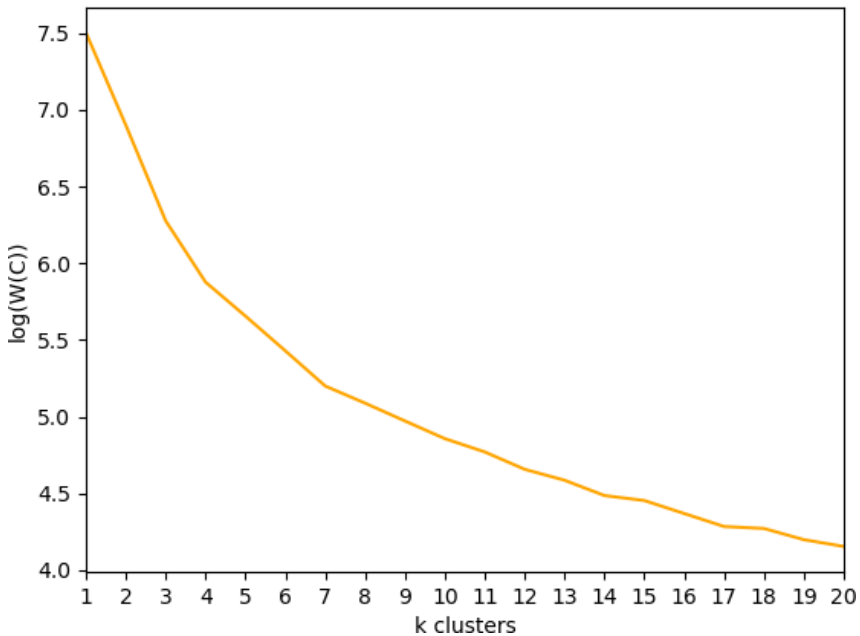
$$W(C) = \sum_{k=1}^{K} N_k \cdot \sum_{C(i)=k} \|\mathbf{x}_i - \boldsymbol{\mu}_k\|^2 \ , \tag{1}$$

  where K is the total number of clusters, $C(i)$ the cluster ID for sample $\mathbf{x}_i$, N_k the number of points in cluster k, and $\boldsymbol{\mu}_k$ the mean of all points in cluster k.
- Also recall that $W(C)$ is a quite natural choice to optimize for compact clusters
- Can we also use $W(C)$ for model selection?
  (the answer will be “yes, but with some additions, which will be the gap statistics”)

Tracking $W(C)$ for Different k

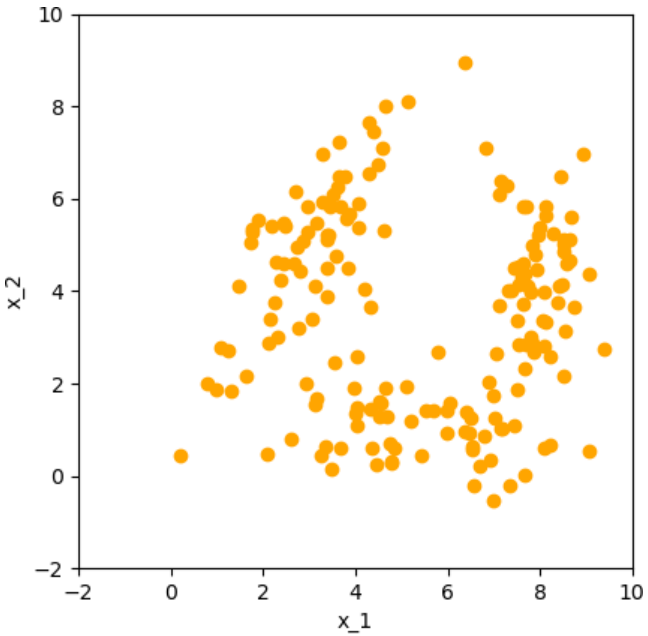
- Investigate the progression of $W(C)$ for different k
- For increasing k, $W(C)$ has to decrease (exceptions are bad local minima):

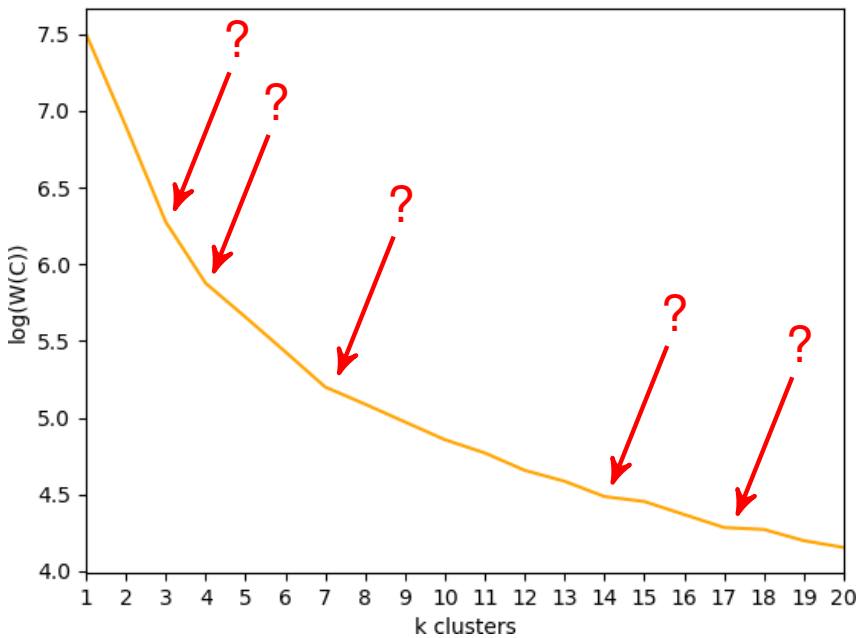




Not-So-Great Possibilities for Model Selection with $W(C)$

- For increasing k, $W(C)$ has to decrease
- Note that $W(C) = 0$ if $k = |X|$ (the trivial solution), hence the optimum k can **not** be found by minimizing $W(C)$
- Also bad is the “elbow method”, to search for a bend on the $W(C)$ curve:
- It is **unclear which bend** is significant. At $k = \{3, 4, 7, 14, 17\}$?





Gap Statistics

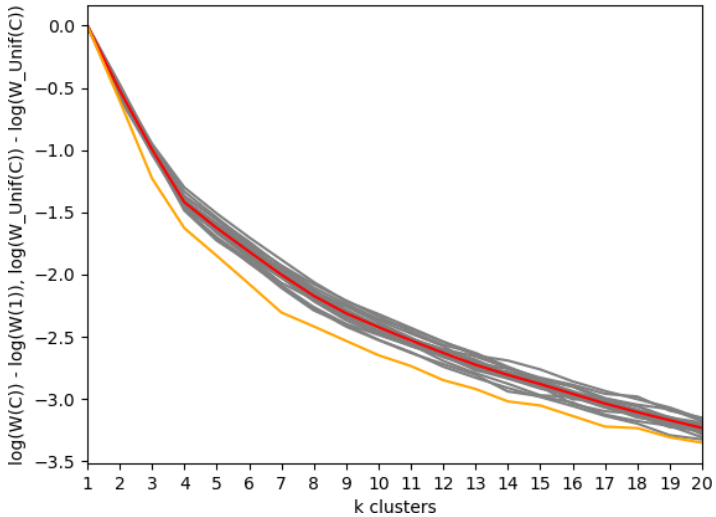
- Tibshirani *et al.* propose to relate $W(C)$ of the actual samples to the $W(C)$ of an artificially created reference
- The reference are samples drawn from the uniform distribution, representing the input with the least possible structure
- Algorithm:
  1. Draw B sets of uniformly distributed samples (Tibshirani uses $B = 20$)
  2. On those distributions, calculate for different k the mean of the log of $W(C)$, denote the result $\log(W_{\text{unif}}(C))$
  3. For k clusters, calculate the gap $G(k)$ as the difference between the reference $\log(W_{\text{unif}}(C)))$ and our log-within cluster distances $\log(W(C))$
  4. Select the optimum k as

$$k^* = \underset{k}{\operatorname{argmin}}\{k | G(k) \geq G(k+1) - s'_{k+1}\} \tag{2}$$

where $s'_{k+1} = s_k \cdot \sqrt{1 + 1/B}$ is an unbiased estimate of the standard deviation s_k of $\log(W_{\text{unif}}(C))$

Within-Cluster Distances on the Uniform Distribution

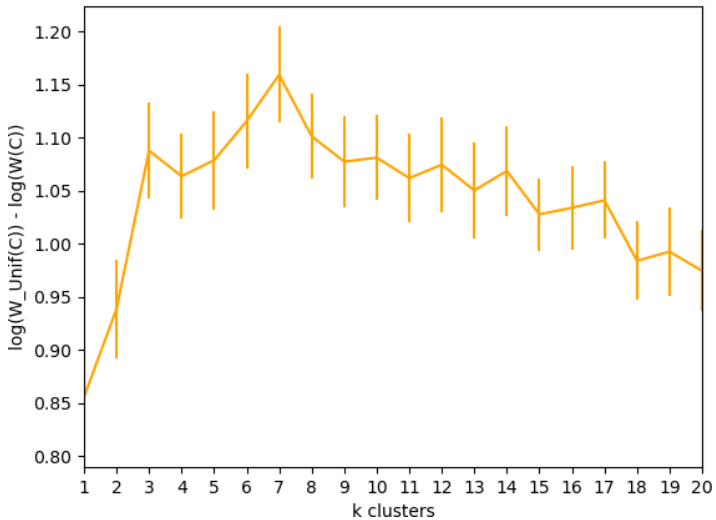
- Gray: $B = 20$ log reference curves from samples drawn from uniform distributions
- Red: Mean $\log(W_{\text{unif}}(C))$
- Orange: $\log(W(C))$ of the actual samples
- (technical remark: all curves are offset-corrected to start at 0)



- Recall that the reference has the least possible structure (uniform distribution)
- Hence, the orange curve can be expected to be lower than the gray curves:
  - Think about Huffman codes: foreseeable events allow more efficient encodings
  - Uniform distributions imply that samples can appear anywhere with identical likelihood, so it will be the “perfect surprise”
  - Conversely, if samples cluster together, they are more likely to be closer to cluster centers. We are less surprised, and we get a lower $\log(W(C))$

Mind the Gap

- Picture on the right: Gaps and standard deviations for the curve differences
- $k^* = 3$ is selected, because the gap at $k = 4$ minus its standard deviation is the first gap that is lower than its predecessor



- Remark: why do the authors choose the first gap with this property (and not the second, third, ...)?
  This is founded in logic, dating back to Ockham's razor: choose the simplest model unless you have a good reason to do otherwise
- Hence, the gap statistics formulates the k-means model selection as "finding the smallest clustering that finds notable structure in the data"

Lecture Pattern Analysis

Part 10: Sampling and GMM MCMC Inference

Christian Riess
IT Security Infrastructures Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg
19. Mai 2025

Introduction

- The model selection task for GMMs particularly has to choose K, the number of mixture components
- GMM model selection is a bit more complicated than for k-means, but it can be done in a fully probabilistic way (i.e., within the same framework)
- Recall that GMM fitting with fixed K has **no closed form** solution, but the **iterative** EM algorithm
- Including the choice of K requires an additional **approximation**, either
  1. Markov Chain Monte Carlo (MCMC) Sampling: approximates the intractable function with a finite number of samples
  2. Variational Inference (VI): approximates the intractable function with a simpler, tractable function
- We start with MCMC, but towards that we cover also the simpler rejection sampling and adaptive rejection sampling[1]

[1] This lecture refers to Bishop Sec. 11–11.1.3, Sec. 11.2.1, Sec. 11.3, and the paper by Rasmussen, which can be found on studOn

Why is Sampling Useful? Example: Evaluation of Expectations

- One application for sampling is to replace analytic calculations, for example:
- Expectations are a backbone of inference, but oftentimes the equation

$$\mathbb{E}[f] = \int f(\mathbf{z}) p(\mathbf{z}) \mathrm{d}\mathbf{z} \tag{1}$$

can not analytically be solved. But if one can sample from $p(\mathbf{z})$, then draw L samples from $p(\mathbf{z})$ and calculate

$$\hat{f} = \frac{1}{L} \sum_{l=1}^{L} f(\mathbf{z}^{(l)}) \tag{2}$$

- This estimator is unbiased, in the sense that $\mathbb{E}[\hat{f}] = \mathbb{E}[f]$ and the variance is

$$\mathrm{var}[\hat{f}] = \frac{1}{L} \mathbb{E}\left[(f - \mathbb{E}[f])^2\right] \tag{3}$$

- Also, the accuracy of the estimator does not depend on the dimension of $\mathbf{z}$, i.e., few samples may suffice

Criticism on the “Lecture-1-Sampler”

- Our sampler from Lecture 1 can operate on general distributions, but its practical usefulness is limited:
- It requires a full representation of the density at every location $\mathbf{x}$
- Hence, whatever our density representation is, it has to be converted to a histogram
- Recall also our conversation in the joint meeting: Histograms are either
  - quite coarse, or
  - quite inefficient (B^D bins), where each bin has to be filled with several data points for sufficient statistics
- Whichever tradeoff we make in the histogram creation propagates as approximation error into the sampler
- Hence, let us look at some alternatives

Sampling from Parametric Standard Distributions

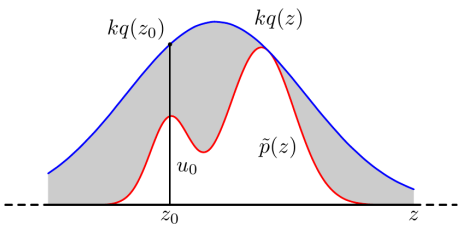
- Analytic mappings from uniform distributions to other distributions exist for
  - Gaussian distributions
  - Exponential distributions $p(y) = \lambda \exp(-\lambda y)$
  - Cauchy distributions $p(y) = \frac{1}{\pi} \frac{1}{1+y^2}$
- This enables a straightforward sampling algorithm:
  - Draw a sample $p(z)$ from a uniform distribution
  - Transform that sample to the target distribution $p(y)$ with the analytic mapping $y = f(z)$
- Note that you need to include the derivative (1-D) or the Jacobian ($>$ 1-D):

$$p(y) = p(z) \left| \frac{\mathrm{d}z}{\mathrm{d}y} \right| \quad p(y_1, \ldots, y_M) = p(z_1, \ldots, z_M) \left| \frac{\partial(z_1, \ldots, z_M)}{\partial(y_1, \ldots, y_M)} \right| \tag{4}$$

- This sampler is highly efficient, since the mapping is analytic
- Conceptually, this is almost identical to our lecture-1-sampler: replace the CDF calculation by the mapping $y = f(z)$

Rejection Sampling

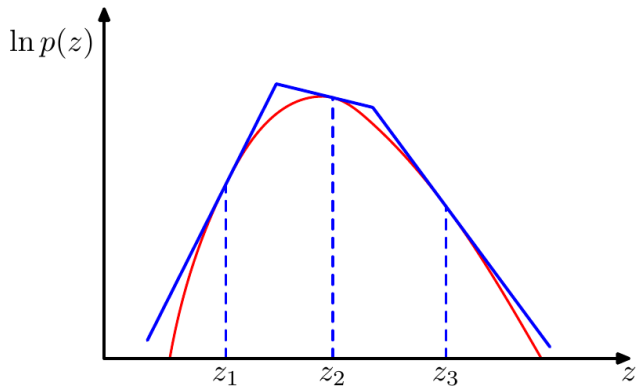
- Distributions are oftentimes more complicated, but it may be possible to obtain $p(z)$ up to an unknown normalization factor Z_P, i.e., $p(z) = \frac{1}{Z_p}\tilde{p}(z)$
- This allows to define a simpler distribution $q(z)$ and a constant k as an envelope to $\tilde{p}(z)$, s.t. $kq(z) \geq \tilde{p}(z)$ for all z
- Then, a sample is obtained in two steps:
  1. Draw sample z_0 from $q(z)$
  2. Draw sample u_0 from a uniform distribution $[0, kq(z_0)]$
  3. Reject (z_0, u_0) if $u_0 > \tilde{p}(z_0)$, otherwise return z_0



- The method is correct, since
  - prior to rejection, the pair (z_0, u_0) is uniformly distributed across the area of the curve $kq(z)$
  - after rejection, the pair (z_0, u_0) is uniformly distributed across the area of the curve $\tilde{p}(z)$

Adaptive Rejection Sampling

- Rejection sampling becomes inefficient if q and p differ too much
- However, a better-fitting envelope q might not have a simple analytic form
- Adaptive Rejection Sampling (ARS) constructs q ad-hoc from $p(z)$
- This works particularly well on log-concave functions, i.e., where derivatives of $\log p(z)$ are non-increasing functions of z



- Fitting a set of lines to the log of the function is equivalent to fitting a piecewise exponential distribution to the original function, i.e.,

$$q(z) = k_i \lambda_i \exp\{-\lambda_i(z - z_i)\} \qquad \hat{z}_{i-1,i} < z \leq \hat{z}_{i,i+1} \tag{5}$$

Sampling in Models with Many Variables or Attributes

- Consider cases where either
  1. a random variable $\mathbf{x}$ has many attributes $\mathbf{x} = (x_1, \ldots, x_d)$, or
  2. a model consists of many dependent random variables $\mathbf{x}_1, \ldots, \mathbf{x}_N$
- The second case applies if when modelling a sampling-based solution for GMM fitting with model selection: $K, \pi_1, \ldots, \pi_K, \mu_1, \ldots, \mu_K, \Sigma_1, \ldots, \Sigma_K$ and some hyperpriors are all random variables

- Let us (somewhat loosely) call such cases “high-dimensional spaces”
- One “sample” is then a full set of assignments to all unknowns
- This makes it almost impossible to use rejection sampling and ARS, because the gap between q and $\tilde{p}$ increases. For example:
  1. Imagine envelopes around all GMM variables
  2. Draw a set of variables, one of them will likely be outside of $\tilde{p}$ (the more variables we have, the more likely this will be)

Markov Chain Monte Carlo Sampling

- Markov Chain Monte Carlo (MCMC) mitigates these issues in high-dimensional spaces
- The idea is to
  - sample from one variable (or attribute) at a time, and
  - to repeat this sampling in an iterative manner using the recently sampled values
- More abstractly, sample in iteration τ from a state space of variables $\mathbf{z}^{(\tau)}$ using the previous iterations $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(\tau-1)}$
- A Markov Chain, specifically, models only first-order statistical dependencies

$$p\left(\mathbf{z}^{(\tau+1)}|\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(\tau)}\right) = p\left(\mathbf{z}^{(\tau+1)}|\mathbf{z}^{(\tau)}\right) \tag{6}$$

- A famous MCMC algorithm is the Metropolis-Hastings method, but we jump right to an important special case, namely Gibbs Sampling

Gibbs Sampling

- We aim to sample from the distribution $p(\mathbf{z}) = p(z_1, \ldots z_M)$ of M random variables (which are somehow initialized)
- Each step of Gibbs sampling updates one variable by drawing from the distribution of that variable conditioned on the others, i.e.,
  1. Initialize $\{z_i : i = 1, \ldots, M\}$
  2. For $\tau = 1, \ldots, T$:
     - 2.1 Sample $z_1^{(\tau+1)} \sim p\left(z_1 | z_2^{(\tau)}, \ldots, z_M^{(\tau)}\right)$
     - 2.2 Sample $z_2^{(\tau+1)} \sim p\left(z_2 | z_1^{(\tau+1)}, z_3^{(\tau)}, \ldots, z_M^{(\tau)}\right)$
     - $\vdots$
     - 2.M Sample $z_M^{(\tau+1)} \sim p\left(z_M | z_1^{(\tau+1)}, \ldots, z_{M-1}^{(\tau+1)}\right)$
- Subsequent samples are correlated (due to the Markov chain)
- However, “the” sample of the complex distribution is a single state after T iterations (i.e., after $M \cdot T$ sub-draws)

Bayesian GMM Fitting: Model Setup

- General ideas for Bayesian modeling:
  - Each GMM parameter $\boldsymbol{\pi}$, $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$ obtains a prior distribution
  - Use conjugate priors: A parameter distribution times its conjugate prior results in the same distribution family as the parameter distribution
  - Parameters to prior distributions (hyperpriors) are usually set to fixed quantities; their specific value is usually not so important
  - Priors shape the distribution in absence of observations
  - Increasing the number of observations "overwrites" the prior
- Bayesian GMM Setup:
  - The conjugate prior for the discrete mixture weights $\boldsymbol{\pi}$ is the Dirichlet distribution $\mathrm{Dir}(\boldsymbol{\pi}|\alpha_1, \ldots, \alpha_M)$ (usually with hyperpriors $\alpha_i = \alpha_0$)
  - The prior for the mean and standard deviation can be written as $p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = p(\boldsymbol{\mu}_k|\boldsymbol{\Sigma}_k) \cdot p(\boldsymbol{\Sigma}_k)$
  - Their conjugate prior is the Gauss-Wishart distribution, where $p(\boldsymbol{\mu}_k|\boldsymbol{\Sigma}_k) = \mathcal{N}(\mathrm{m}_0, \beta_0\Sigma_k)$ and $p(\boldsymbol{\Sigma}_k) = \mathrm{Wish}(\boldsymbol{\Sigma}_k|\nu, \mathbf{V})$ with hyperpriors m_0, β_0, ν, $\mathbf{V}$.

Gibbs Sampling Applied to GMM Models (1/2)

- Define a conditional distribution and a prior for each GMM variable $\boldsymbol{\pi}$, $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$, the hidden responsibilities $\mathbf{Z}$[2].
- Without going too much into detail: each individual distribution is chosen such that it is easy to sample from it (e.g., a normal/Gamma/Dirichlet distribution)
- Iteratively sample from each distribution. Responsibility are sampled for $K + 1$ components, which enables the creation of new clusters
- Repeat this until the number of components somewhat stabilizes in an interval
- When stopping the iteration, the current state is one sample, i.e., it is one specific GMM with a specific number of components and parameters
- Fig. 2 (right) in Rasmussen's paper shows that the sampled GMMs vary in size between 15 and 25 components

[2] Browse Rasmussen's paper if you want to know more!

Results by Rasmussen

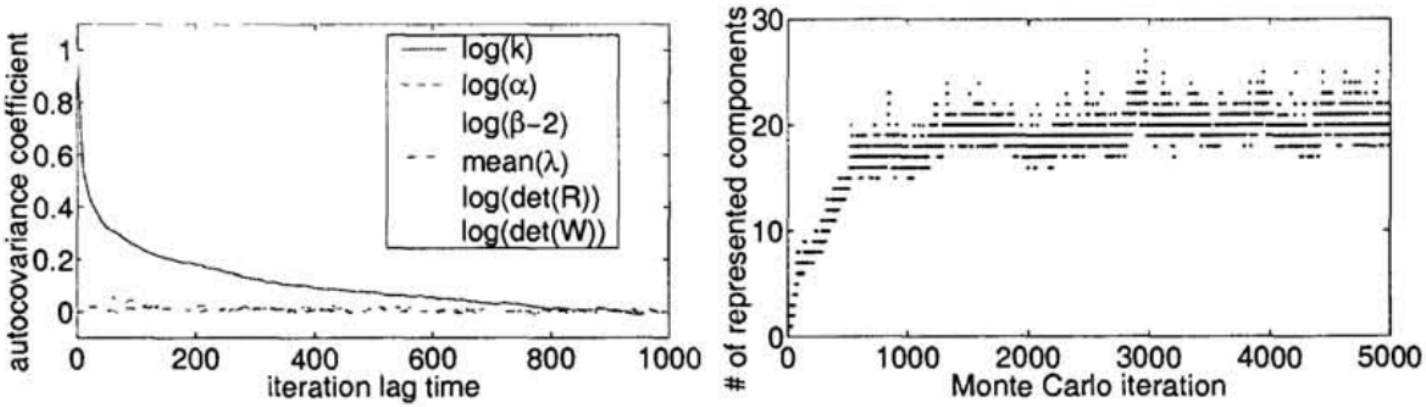


Figure 2: The left plot shows the auto-covariance length for various parameters in the Markov Chain, based on 10^5 iterations. Only the number of represented classes, k_{rep}, has a significant correlation; the effective correlation length is approximately 270, computed as the sum of covariance coefficients between lag -1000 and 1000. The right hand plot shows the number of represented classes growing during the initial phase of sampling. The initial 3000 iterations are discarded.

Remarks

- Note that “one sample” may be obtained from multiple draws:
  - Lecture-1-sampler: one draw
  - Analytic Samplers: one draw
  - Rejection Sampling / ARS: two draws (times one plus number of rejects)
  - MCMC Sampling: $M \cdot T$ draws
- MCMC Sampling is quite popular, because modelling the individual conditionals and priors is relatively straightforward
- Hence, MCMC Sampling has high runtime cost but little “thinking cost”