STCS 6701: Foundations of Graphical Models: Homework 1

Maxime TCHIBOZO (MT3390)

Monday October 12, 2020

1 Gibbs Sampling for Multivariate GMM

Chart patterns are a commonly-used tool in the analysis of financial data. In this homework, I developed Gaussian Mixture Models using Gibbs Sampling to cluster stock market Time Series according to their adjusted close price over time. Each Time Series in the dataset corresponds to a vector containing the adjusted close price of a stock over 50 days (Appendix: Figure 1).

We fix the number of clusters to $K = 4$, and use a set of $N = 200$ Time Series. Each Time Series assigned to cluster k is assumed to be drawn from a Multivariate Gaussian with a 50x1 cluster mean $\mu^{(k)}$ and 50x50 covariance matrix $\Sigma^{(k)}$. All parameters are learnt through Gibbs sampling (cluster proportions $\Theta \sim$ Dirichlet; cluster assignments $Z_i \sim$ Multinomial; cluster means $\mu^{(k)} \sim$ Multivariate Normal; cluster covariance $\Sigma^{(k)} \sim$ Inverse Wishart) (See Gibbs_GMM_multivariate.ipynb file for the exact mathematical derivations).

Initialization and Convergence of the Gibbs Sampler:

To evaluate the "burn-in" time necessary for the Gibbs sampler to converge, we plot the log joint likelihood of the model at each iteration (Appendix: Figure 3), using the conditional probabilities and priors. Running 7 chains with randomly initialized parameters (randomness is generated by the prior distributions) for 1000 iteration shows that convergence begins after 300-400 iterations for the log joint likelihood, and at the same time for the cluster proportions (Appendix: Figure 4). Since the model for the Data Generating Process is high-dimensional, sampling takes time, and a trade-off must be established between the length of the burn-in phase, and the number of samples we expect to draw from the converged posteriors. We set a burn-in limit to start sampling after 300 iterations.

Statistical Properties of the Sampled Parameters:

Once in the sampling phase, we sample the parameters (Z, μ, Σ, Θ) every 10 iterations. In theory, MCMC samples are independent of each other if samples are sufficiently spaced out in time. Here time constraints combined with the need to collect 1000 samples at least for statistical significance implies that the 10 iteration window is likely too small for true independence between samples. We study the quality of our samples in the using N_{eff} (effective sample size) and Monte Carlo Error.

We measure statistics of the sampled posterior parameters for 3 separate chains of 1000 samples. For each chain, we measure cluster probabilities (parameters of the multinomial), the first dimension of each cluster mean $\mu_0^{(k)}$, the first dimension of each cluster covariance matrix $\Sigma_{00}^{(k)}$, and the number of Z_i assignments to each cluster (Appendix: Table 2; Table 3; Table 4; Table 5). While the tables indicate the Θ and Z parameters are sampled from a distribution, the mean and covariance parameters seem to be fixed point estimates, meaning after convergence of the sampler these parameters no longer change. This is a first indication of the limitations of the model in representing the data.

We use effective sample measure N_{eff} as a tool to confirm this intuition. N_{eff} indicates that the 1000 topic proportions allow us to generate approximately 133 independent data points (we have initially $N = 200$ points). For the other parameters, N_{eff} is either less than 1 or undefined, meaning these samples cannot be considered independent of each other.

The Monte Carlo Error (Appendix: Table 1) over the 3 separate chains (which compares within chain variance to across chain variance) indicates that the chains are likely independent of each other with respect to some Θ, μ, Σ parameters, with the exception of $\Sigma_{00}^{\{1,2\}}$.

Predictions and Model Behavior:

To check whether the model accurately captures the properties of Time Series, we use sampled parameters of each cluster to generate new "synthetic" samples (Appendix: Figure 2). Data from clusters 0,1 and 3 are realistic, but cluster 2 is not: the y-scale has a magnitude of 10^{-4}, whereas the true data is standardized and takes values of magnitude 10^0.

Another way to check for model behavior is to directly visualize cluster assignments (Appendix: Figure 5; Figure 6). We select the assignments from sample 0 of chain 0 for cluster 1 (which contained only 3 data points), and cluster 3 (which contained 15 data points). Cluster 1 seems to capture Time Series which have sharp day-on-day increases, whereas Cluster 3 seems to capture Time Series which have longer term increasing or decreasing trends.

Limitations of the model

This GMM model is limited for several reasons. Firstly, stock market data is

notoriously noisy and random, extracting meaningful clusters from such data requires additional preprocessing, notably to realign the Time Series. Secondly, our analysis found that some model parameters (Σ, μ) were not drawn from distributions, but rather stayed fixed during sampling. This could be due to the small size of the dataset ($N = 200$), but is also likely due to the high dimensionality of the Inverse-Wishart and Multivariate Normal distributions. These distributions also routinely cause numerical errors due to the low likelihood of sampled parameters, to domain support constraints (Σ must be positive semi definite), and to the possibility of some clusters being assigned no points (cluster probability imbalance). While we chose uninformative priors in this approach, we could have fixed the covariance matrices for each cluster to simplify the learning process.

2 Final Project Brainstorming

My final project will analyze public and private recruitment data (resumes, job postings, candidate data). The main idea of the project will be to evaluate how candidates from different demographics (race, gender, age) interact differently to job postings. One outcome could be to measure whether certain words in job postings deter candidates from applying to jobs.

(a) Variables in the data.

The public data contains resumes of candidates in free text format, as well as some categorical variables (age, location, gender). The private data contains data from minority software engineers who applied to several companies. This data includes information on race, quantitative assessment scores, jobs applied to and companies. These variables are highly correlated, and an outcome for the project could be to evaluate which variables can be considered conditionally independent.

(b) Latent variables

We can introduce latent variables which capture candidates' quantitative skills, personality, and how well candidates match a specific job posting.

(c) Research questions

- Can we identify words most indicative of racial or gender bias in text documents (similar to sentiment analysis)?
- Can we determine a behavioral fit scoring or matching metric for candidates?
- Can we create a recommendation system which combines outcomes (admitted/rejected) with user metadata in free text form (word embeddings such as Word2Vec or BERT)?

3 Appendix: Tables, Figures for Problem 1

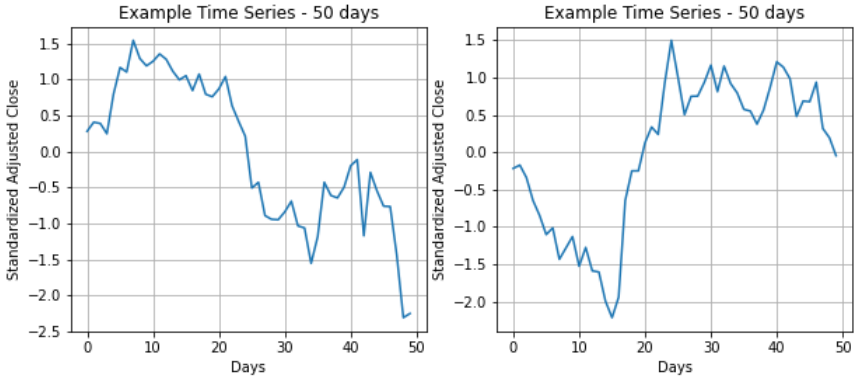

Figure 1: Example of "True" 50-day Time Series in the dataset

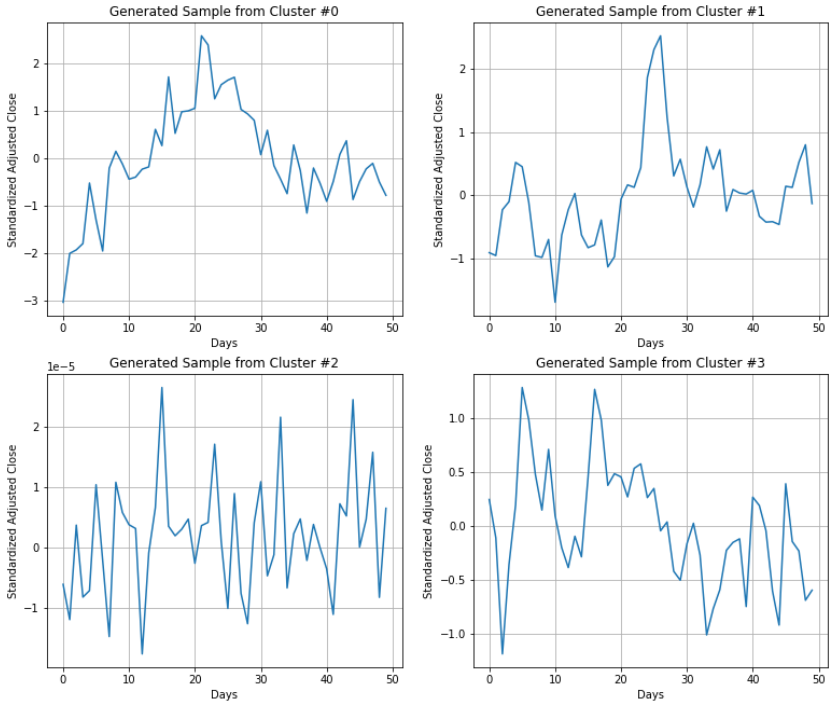

Figure 2: "Synthetic" 50-day Time Series Generated from Posterior-sampled parameters

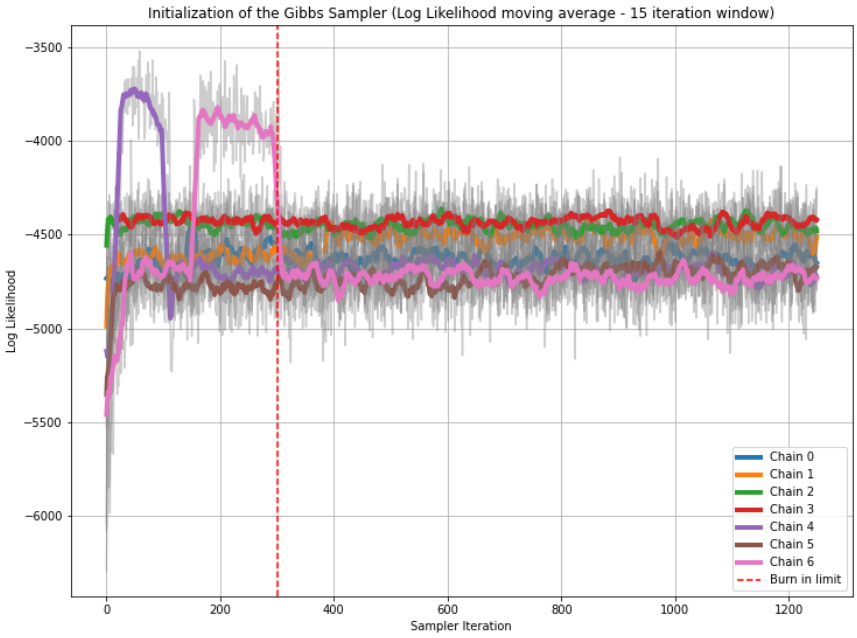

Figure 3: Initialization and Convergence of the Gibbs sampler: Log-likelihood

Parameter	Monte Carlo Error - 3 chains
Θ_0	0.001675
Θ_1	0.290385
Θ_2	0.287756
Θ_3	0.001601
$\mu_0^{(0)}$	0.418118
$\mu_0^{(1)}$	0.039312
$\mu_0^{(2)}$	0.395128
$\mu_0^{(3)}$	0.277363
$\Sigma_{00}^{(0)}$	0.016051
$\Sigma_{00}^{(1)}$	1.197761
$\Sigma_{00}^{(2)}$	1.641796
$\Sigma_{00}^{(3)}$	0.036228

Table 1: Monte Carlo Error over the 3 chains (1000 samples each)

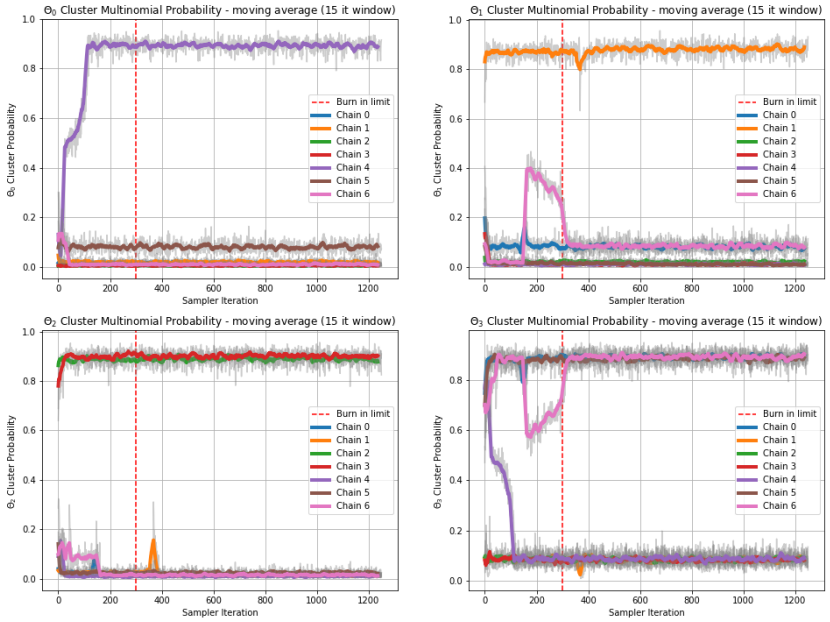

Figure 4: Initialization and Convergence of the Gibbs sampler: Cluster Proportions Θ

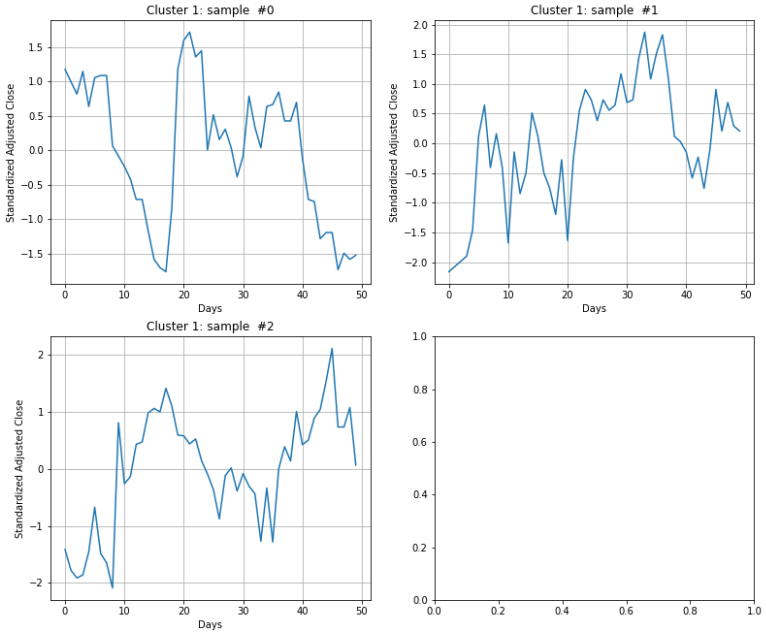

Figure 5: Sampler Cluster Assignments: Cluster 1

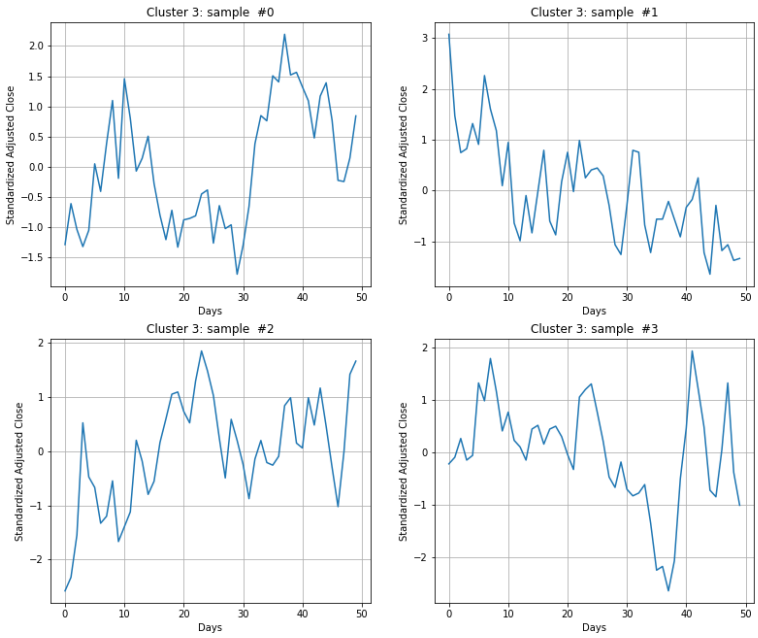

Figure 6: Sampler Cluster Assignments: Cluster 3

Parameter	mean	sd	2.5%	25%	50%	75%	97.5%	N_{eff}
Θ_0	0.015	8.015e-03	0.003	0.009	0.014	0.019	0.035	133.333
Θ_1	0.025	2.637e-02	0.005	0.013	0.019	0.027	0.099	133.333
Θ_2	0.875	3.854e-02	0.768	0.862	0.880	0.899	0.922	133.333
Θ_3	0.085	2.641e-02	0.046	0.067	0.083	0.100	0.136	133.333
$\mu_0^{(0)}$	-1.421	4.441e-16	-1.421	-1.421	-1.421	-1.421	-1.421	0.399
$\mu_0^{(1)}$	-0.141	0.000e+00	-0.141	-0.141	-0.141	-0.141	-0.141	NaN
$\mu_0^{(2)}$	-0.254	5.551e-17	-0.254	-0.254	-0.254	-0.254	-0.254	0.399
$\mu_0^{(3)}$	-0.231	5.551e-17	-0.231	-0.231	-0.231	-0.231	-0.231	0.399
$\Sigma_{00}^{(0)}$	0.029	1.041e-17	0.029	0.0294	0.029	0.029	0.029	0.399
$\Sigma_{00}^{(1)}$	0.028	3.469e-18	0.028	0.028	0.028	0.028	0.028	0.399
$\Sigma_{00}^{(2)}$	4.967	8.882e-16	4.967	4.967	4.967	4.967	4.967	0.399
$\Sigma_{00}^{(3)}$	0.496	0.000e+00	0.496	0.496	0.496	0.496	0.496	NaN

Table 2: Sample Statistics for chain #0 (1000 samples)

Parameter	mean	sd	2.5%	25%	50%	75%	97.5%	N_{eff}
Θ_0	0.010	6.928e-03	0.001	0.005	0.008	0.013	0.027	133.333
Θ_1	0.891	2.409e-02	0.838	0.876	0.892	0.908	0.933	133.333
Θ_2	0.019	9.730e-03	0.006	0.012	0.018	0.025	0.043	133.333
Θ_3	0.080	2.188e-02	0.042	0.065	0.078	0.094	0.127	133.333
$\mu_0^{(0)}$	-1.395	2.220e-16	-1.395	-1.395	-1.395	-1.395	-1.395	0.399
$\mu_0^{(1)}$	-0.009	0.000e+00	-0.009	-0.009	-0.009	-0.009	-0.009	NaN
$\mu_0^{(2)}$	1.038	0.000e+00	1.038	1.038	1.038	1.038	1.039	NaN
$\mu_0^{(3)}$	-1.178	2.220e-16	-1.178	-1.178	-1.178	-1.178	-1.178	0.399
$\Sigma_{00}^{(0)}$	0.017	0.000e+00	0.017	0.017	0.017	0.017	0.017	NaN
$\Sigma_{00}^{(1)}$	3.234	8.882e-16	3.234	3.234	3.234	3.234	3.234	0.399
$\Sigma_{00}^{(2)}$	0.0847	0.000e+00	0.085	0.085	0.085	0.085	0.085	NaN
$\Sigma_{00}^{(3)}$	0.390	5.551e-17	0.390	0.3902	0.390	0.390	0.390	0.399

Table 3: Sample Statistics for chain #1 (1000 samples)

Parameter	mean	sd	2.5%	25%	50%	75%	97.5%	N_{eff}
Θ_0	9.75e-3	6.99e-3	1.40e-3	4.84e-3	7.93e-3	1.29e-2	2.80e-2	133.33
Θ_1	9.01e-1	2.27e-2	8.52e-1	8.87e-1	9.02e-1	9.17e-1	9.42e-1	133.33
Θ_2	4.80e-3	5.04e-3	1.11e-4	1.26e-3	3.16e-3	6.84e-3	1.89e-2	133.33
Θ_3	8.46e-2	2.14e-2	4.72e-2	6.94e-2	8.27e-2	9.71e-2	1.31e-1	133.33
$\mu_0^{(0)}$	-2.66	≈ 0	-2.66	-2.66	-2.66	-2.66	-2.66	0.399
$\mu_0^{(1)}$	-4.77e-2	≈ 0	-4.77e-2	-4.77e-2	-4.77e-2	-4.77e-2	-4.77e-2	0.399
$\mu_0^{(2)}$	-8.80e-7	≈ 0	-8.80e-7	-8.80e-7	-8.80e-7	-8.80e-7	-8.80e-7	0.399
$\mu_0^{(3)}$	-5.67e-1	≈ 0	-5.67e-1	-5.67e-1	-5.67e-1	-5.67e-1	-5.67e-1	NaN
$\Sigma_{00}^{(0)}$	7.05e-2	≈ 0	7.05e-2	7.05e-2	7.05e-2	7.05e-2	7.05e-2	NaN
$\Sigma_{00}^{(1)}$	3.91	≈ 0	3.91	3.91	3.91	3.91	3.91	0.399
$\Sigma_{00}^{(2)}$	≈ 0	≈ 0	≈ 0	≈ 0	≈ 0	≈ 0	≈ 0	NaN
$\Sigma_{00}^{(3)}$	5.02e-1	≈ 0	5.02e-1	5.02e-1	5.02e-1	5.02e-1	5.02e-1	0.399

Table 4: Sample Statistics for chain #2 (1000 samples)

	mean	sd	2.5% percentile	25%	50%	75%	97.5%
Z = 0	177.581	6.020	157.975	177.0	179.0	180.0	182.0
Z = 1	9.584	8.274	2.000	3.0	3.0	16.0	25.0
Z = 2	8.685	7.476	2.000	2.0	2.0	16.0	20.0
Z = 3	4.150	4.213	2.000	2.0	3.0	3.0	17.0

Table 5: Cluster Assignment counts (Z) Statistics for chain #0 (1000 samples)