Problem set 2: Units 4 & 5

CS146

Minerva University

Professor Volkan

April 19, 2024

Report

Task Description: In this problem set, we revisit the polynomial linear regression models from Session 10. We fit polynomials of various degrees to the data set below. You might find it helpful to revisit the pre-class and breakout workbooks from that class to review the work we did there. The data set used here is different from the ones used in class.

Data Set Exploration:

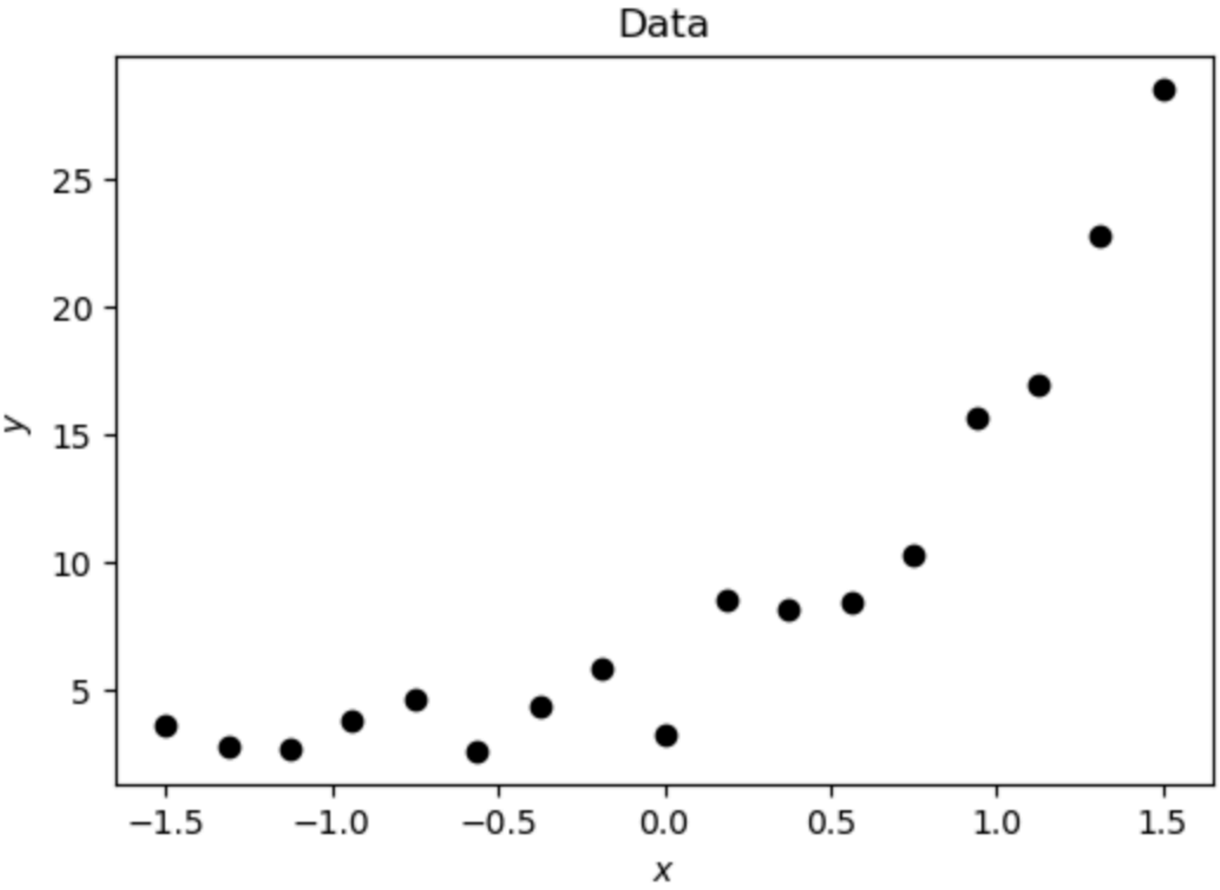


Figure 1. The data set used for the given assignment.

The polynomial models all have a likelihood function of the form

$$y_1 \sim Normal(mu_i, \ sigma^2)$$

$$mu_i = \ a \ + \ b_1 x_1 + b_2 x^2_i + \ ... \ + \ b_k x^k_i$$

where **k** is the degree of the polynomial.

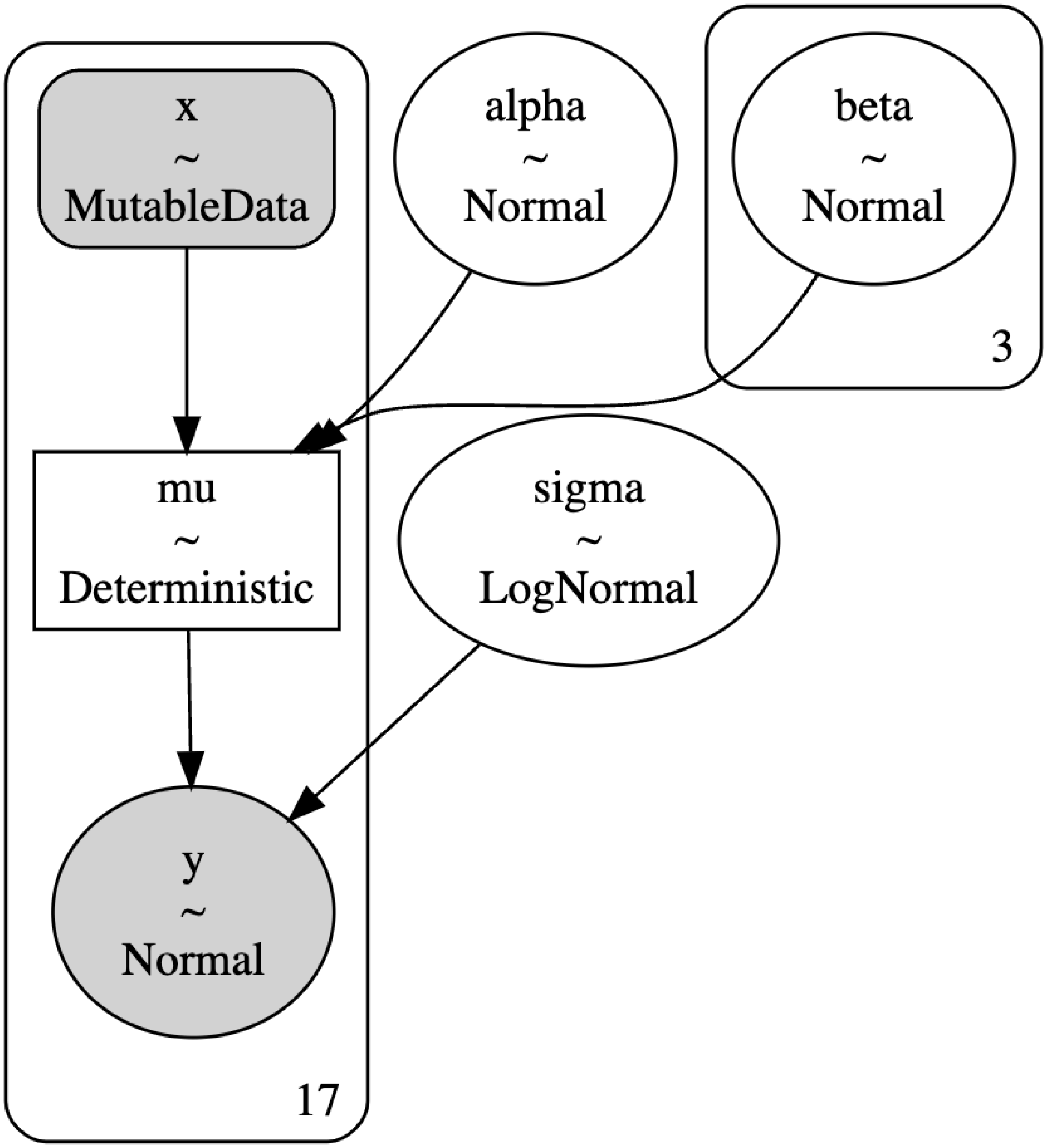


Figure 2. The polynomial model is provided in the problem set.

Model Evidence Estimation

The given section explores why it is possible to use function-fitting algorithms to estimate the model evidence by providing conceptual and mathematical explanations.

Conceptual Explanation

The reason why it is possible to use function-fitting algorithms like ADVI or Laplace Approximation to compute model evidence is that these algorithms can estimate the best fit for the model given the set of data. Overall, the function fitting algorithms find the best-fitting function that describes the provided data and then utilize the same function to estimate the model evidence. We can utilize Bayes' theorem in order to explain this reasoning in greater detail. Conceptually, these algorithms consider the probability of data given the defined model, which is the numerator in Bayes' theorem, while accounting for the probability of the model itself, which is the denominator or a marginal likelihood of the theorem. Therefore, by accounting for relative probabilities of the data under various models, algorithms can identify the best-fit model or, as we call it, identify the highest model evidence. Moreover, since we are working with probabilities, we have to ensure that these probabilities sum up to 1. The denominator or model evidence takes care of it by playing the role of a normalizing constant.

Mathematical Explanation

We can reference the ADVI algorithm to explain how model evidence is relevant. In ADVI, model evidence shows up as a term that we want to minimize in the objective function. This term is the negative log of the marginal likelihood and is proportional to model evidence. As a result, when the algorithm seeks to minimize this term (which is the main goal), it ends up maximizing the model evidence. To explain the relevance of the model evidence from the perspective of the Laplace Approximation, it is important to acknowledge that the model evidence is given by the normalizing constant of the approximate posterior distribution, which can be computed using the parameters found by the fitting algorithm. In both cases, the model evidence is a crucial part of the function fitting algorithm and is used to evaluate the relative fit of different models to the data. Finally, we say that mathematically the model evidence is expressed as an integral of the product of the likelihood function and the prior probability over all possible parameter values. When it comes to ADVI, this integral is approximated through a variational distribution. Because of this we can calculate the model evidence efficiently. In the Laplace Approximation, we approximate the model evidence by computing the product of the maximum likelihood and the determinant of the Hessian matrix. This results in the approximation of the curvature of the log-likelihood function at the maximum likelihood point.

Laplace Approximation

The Laplace approximation involves approximating a target distribution by fitting a Gaussian distribution, leveraging the Central Limit Theorem's assertion that, for sufficiently large sample sizes, sample means tend to follow a Normal distribution. By extending this logic, we approximate the mean of the sample using a Normal distribution. To fit a Normal distribution to a dataset, we start by finding the log posterior probability density function (pdf), which is differentiable at all points. Employing an optimization procedure, we locate the point where the derivative equals zero, indicating the peak of the distribution. This peak serves as the center of our Normal distribution, around which we model the remainder of the distribution. Utilizing second derivatives of the log posterior, we construct a covariance matrix (Hessian matrix) to shape the Normal distribution based on the dimensions of the final posterior. When fitting marginal distributions instead of the joint final distribution, we can choose between two approaches: Full-rank approximation, which considers covariance between all dimensions, and mean-field approximation, which assumes either no covariance or that it has no impact on the results. The latter option significantly reduces computational complexity, as the Hessian matrix scales quadratically with each additional dimension ($O(n^2)$). Once the model is fitted, we determine the normalization constant of the distribution. This constant serves as our estimate of the model evidence, enabling comparisons between different models. It represents the probability of the data occurring given a specific model, also known as the marginal likelihood in Bayes Theorem ($P(D \mid M_i)$).

Comparing ADVI and Laplace Approximation (optional)

This section focuses on comparing both algorithms.

ADVI (Automatic Differentiation Variational Inference)

Computation: ADVI utilizes the Evidence Lower Bound (ELBO) to estimate the model evidence. The ELBO serves as a lower bound for the log of the model evidence and aims to minimize the Kullback-Leibler (KL) divergence between the approximating distribution q and the target distribution p. Maximizing the ELBO indirectly minimizes this divergence by optimizing a function involving the expected log-likelihood of the data under the variational distribution and the entropy of the variational distribution itself. Typically, a standard distribution like the Normal is chosen as the variational distribution to simplify calculations and sampling. Approximating the expected log-likelihood of the data under the variational distribution involves sampling methods due to its complexity. Stochastic optimization techniques, such as the Adam algorithm, are employed to efficiently maximize the ELBO, thereby bringing the approximation closer to the true model evidence.

Pros: ADVI efficiently handles large datasets and complex models, often converging faster than traditional MCMC methods.

Cons: The approximation may lack accuracy, particularly for posteriors that are poorly represented by the chosen simpler distribution. Since the model evidence is not directly computed, there may be inaccuracies in its estimation.

Laplace Approximation

Computation: The Laplace approximation approximates the posterior distribution around its mode using a Normal distribution. This approximation relies on the curvature around the mode, as determined by the Hessian matrix. The estimation of model evidence is indirect, relying on evaluating the fit of this approximation to the true posterior.

Pros: It is simple and straightforward to implement with minimal parameters. Laplace approximation tends to be accurate for unimodal posteriors that are approximately normal and not heavily skewed.

Cons: In high-dimensional models, computational costs increase due to the need to calculate the Hessian matrix. It may provide a poor approximation for multimodal posteriors or those that deviate significantly from normality.

Mean-Field vs Full Rank ADVI

The given section will focus on comparing the performance of mean-field vs full-fank ADVI algorithms.

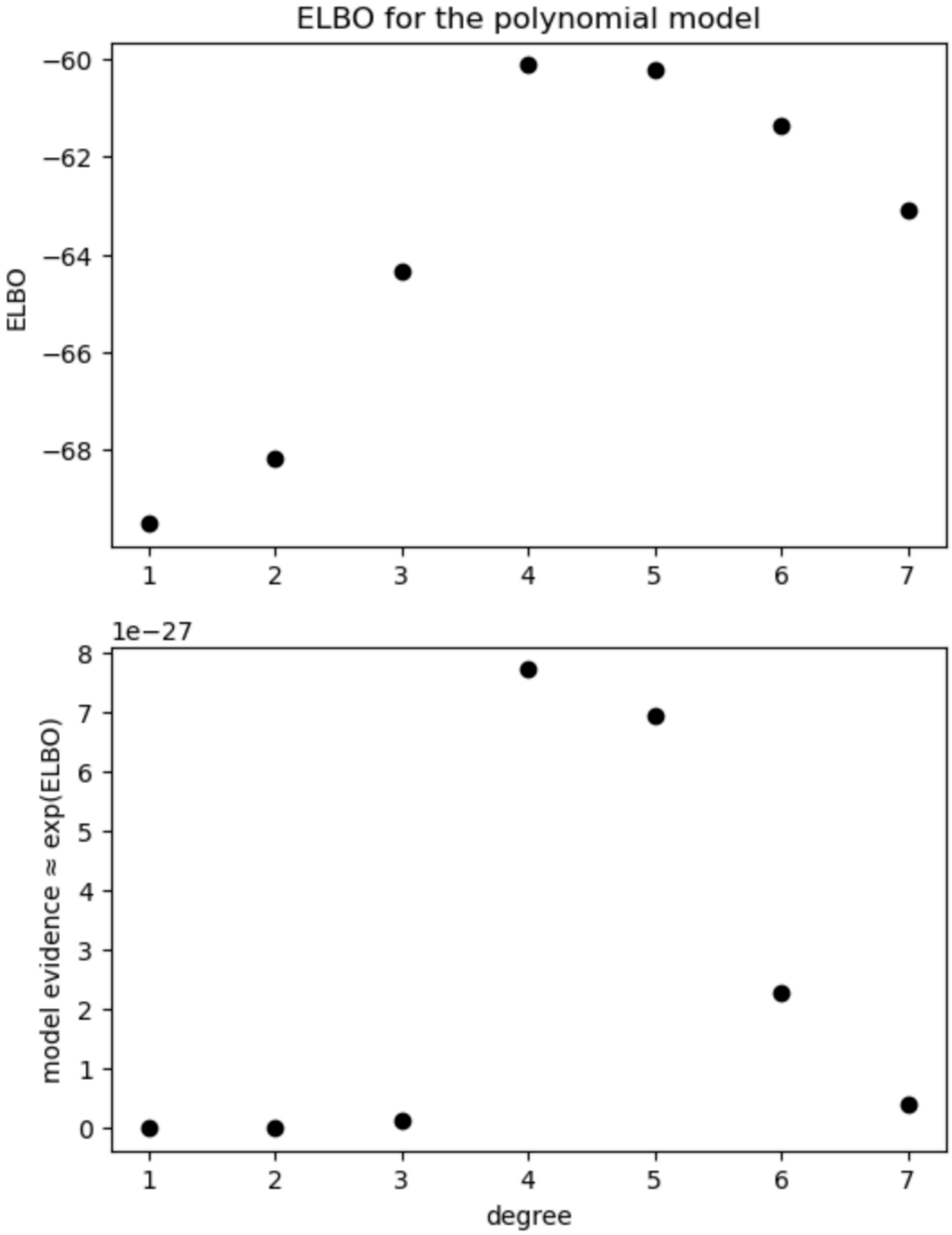


Figure 3. The ELBO values and Model Evidence are computed by the Mean-Field Algorithm.

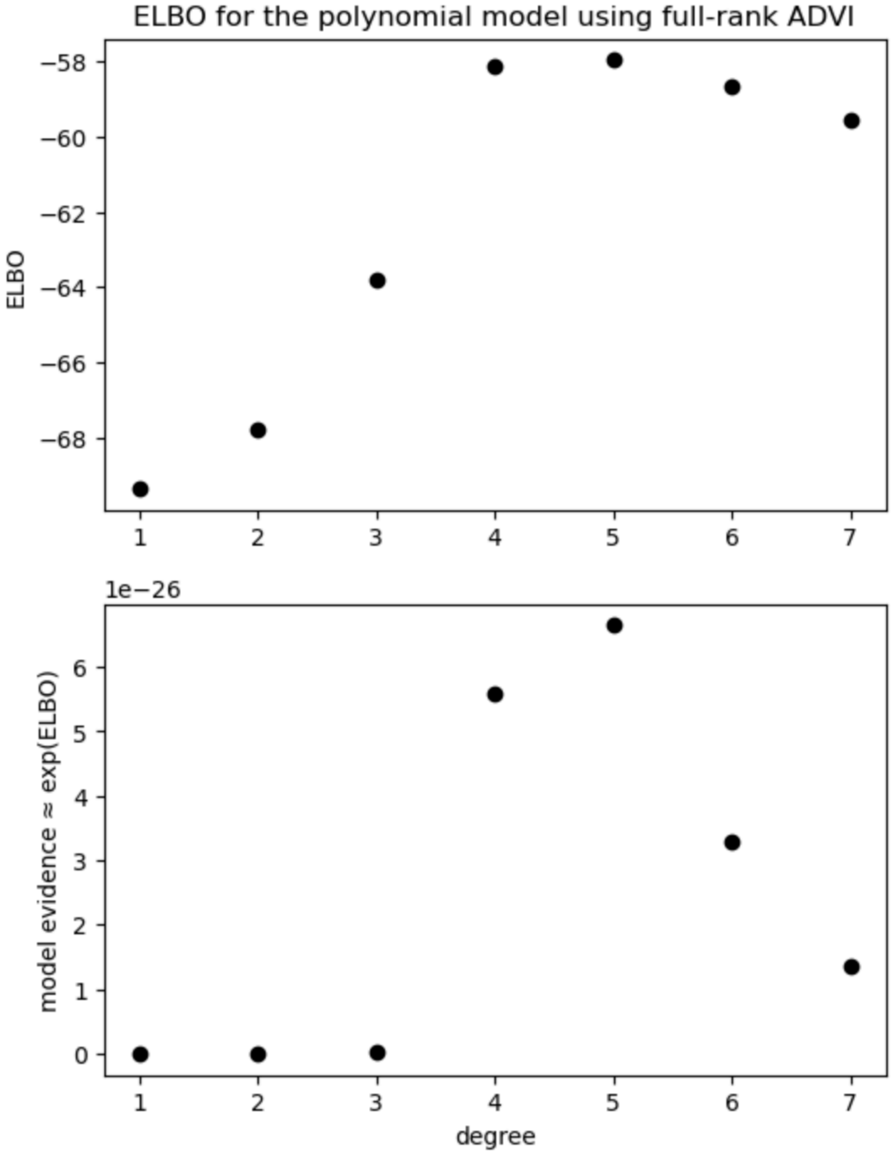


Figure 4. The ELBO values and Model Evidence are computed by the Full Rank ADVI ALgorithm.

How much worse is the mean-field approximation than the full-rank approximation? Quantify your answer to this question and explain your work.

In order to quantify and compare the difference to assess the efficacy of each algorithm, we can compare the evidence lower bound (ELBO) values obtained for each polynomial degree after using both algorithms. The Evidence Lower Bound (ELBO) is useful here because it approximates the logarithm of the model evidence, indicating the quality of the posterior approximation. By comparing ELBO values, we can determine which approximation fits the data better. We'll calculate differences in ELBO values for each degree, with positive differences suggesting a better fit by the full-rank approximation. Additionally, computing the average difference across all degrees provides an overall performance comparison. A plot of both ELBO values can reveal if one method consistently outperforms the other or if performance varies with model complexity.

Average ELBO differences: 1.6399824784553831

ELBO differences between full-rank ADVI and mean-field ADVI: [0.14139787 0.39194297 0.50889816 1.98200892 2.25974017 2.67563479 3.52025446]

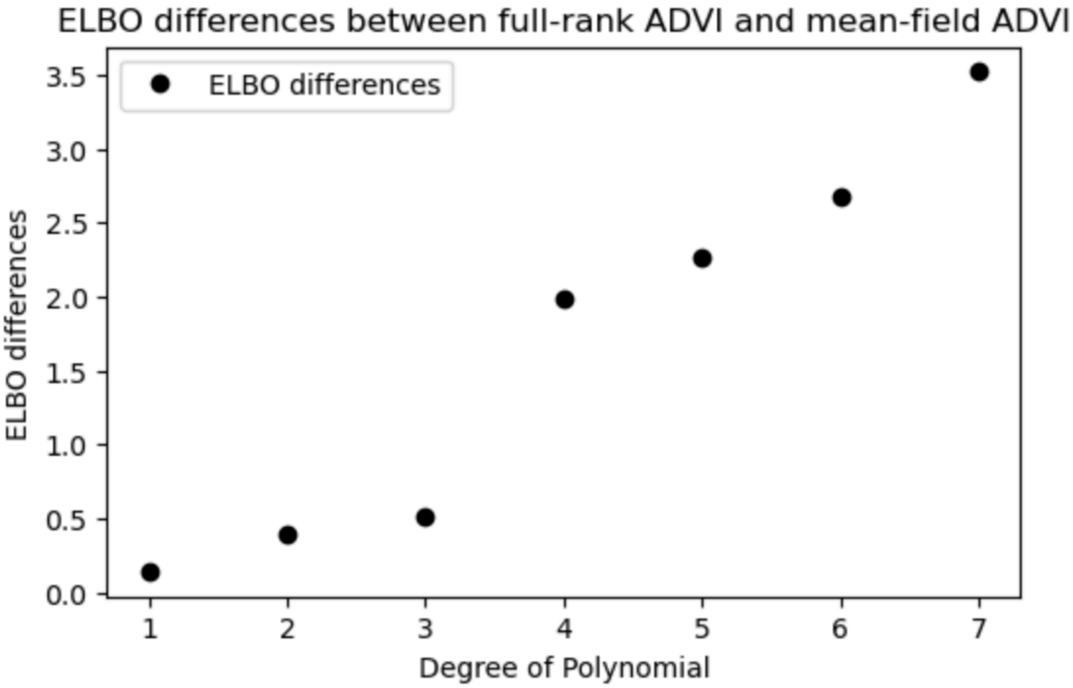


Figure 5. ELBO Differences between full-rank and mean-field algorithms.

Firstly, we calculated the difference between ELBO values per each polynomial along with the average difference between the two algorithms. By observing the resulting values and the plot above, we can conclude that full-rank approximation provides better estimates compared to the mean field because we can observe a positive and increasing trend in ELBO values as the degree of polynomials increases.

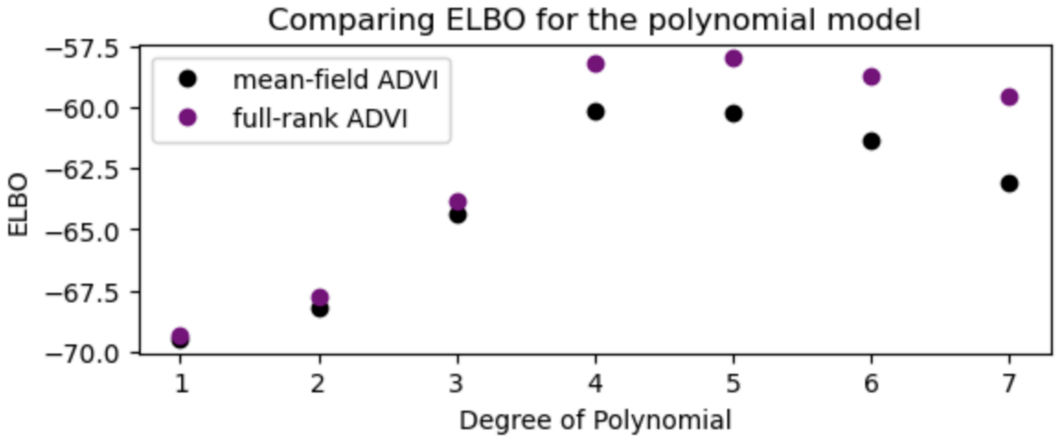
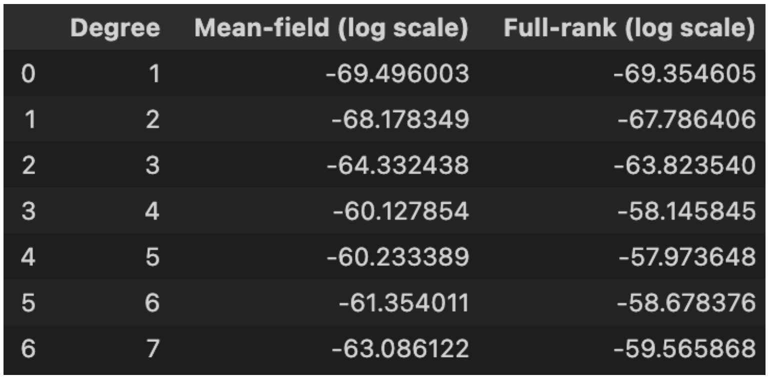


Figure 6. Comparison of full-rank and mean-field algorithms' ELBO Values.

The given plot showcases ELBO values for each approximation across polynomial degrees. This plot reveals that both methods have similar trend in their ELBO values per polynomial; however, full-rank shows a higher values per each polynomial compared to the mean-field. The average ELBO difference supports the previous statement of similar trends in their computed values because the average ELBO difference is only 1.6399.



	Degree	Mean-field (log scale)	Full-rank (log scale)
0	1	-69.496003	-69.354605
1	2	-68.178349	-67.786406
2	3	-64.332438	-63.823540
3	4	-60.127854	-58.145845
4	5	-60.233389	-57.973648
5	6	-61.354011	-58.678376
6	7	-63.086122	-59.565868

Table 1. Log Scale Values for full-rank and mean-field algorithms.

The table above shows exact log scale values. The highest ELBO values for the mean-field are situated around the 4th degree (-60.127), while the full rank's highest value is around the 5th polynomial (-57.973). Given this difference, it is always important to consider trade-offs between using either of the models. Since mean-field does not perform fully incorrectly, we might rely on it due to its lower computational cost. However, if we need precision, full-rank is more appropriate.

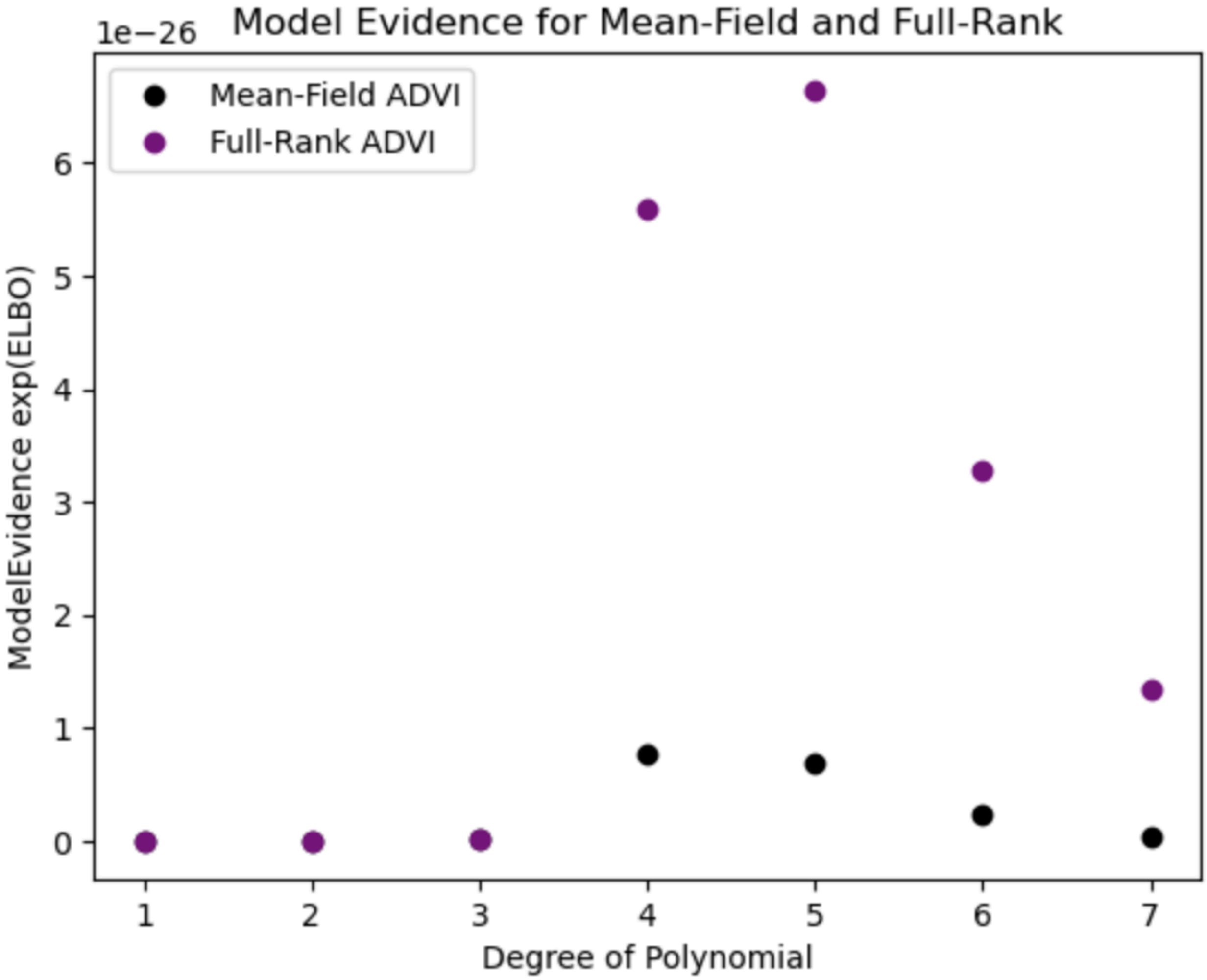


Figure 7. Comparison of model evidence of full-rank and mean-field algorithms

Do any of our conclusions change about which degrees are better or worse than others?

Model evidence (approximated by exp(ELBO)) can also be a useful tool for comparisons. The plot above shows that depending on the algorithm, the best polynomial changes from degree 4 with a mean field to degree 5 with a full-rank algorithm.

Conclusion

In conclusion, the comparison between the mean-field and full-rank ADVI algorithms reveals differences in their performance metrics. While both methods exhibit decreasing ELBO values with increasing polynomial degree, suggesting improved model fits with higher complexity, the full-rank ADVI consistently yields slightly higher ELBO values, indicating better optimization and closer alignment with the data. Additionally, the full-rank ADVI generally produces higher model evidence values compared to the mean-field approximation, albeit with small numerical values that require cautious interpretation. Therefore, we can conclude that the full-rank ADVI algorithm demonstrates better performance in terms of both ELBO and model evidence across various polynomial degrees. Moreover, we can see how the "best-fit polynomial" changes depending on the algorithm as shown in the plot above. The earlier comments about trade-offs are highly relevant to ensure the accuracy of the model.

References

Sessions 10, 22, 23 from CS146 course.

AI Statement

The models, analysis, and other substantial part of the assignment is self-completed.

Video

https://www.loom.com/share/3cd3eb4254364148887406490c4b3027?sid=76ce8f5e-be4c-460b-a60f-112e7356ce28