ISyE 6416 - Spring 2023
Final Project Report

Team Member Names: Alp Serdaroglu

Project Title: Estimating the Bias in the Election Polls for 2023 Turkish General Election

Contents

1 Problem Statement

Elections are an essential part of democratic life. They ensure that people are represented and participate in their countries' decision-making. Through elections, citizens can vote who will represent their views in the parliament and oversee the activities of the government. Moreover, elections provide citizens with an opportunity to participate in the political process and have a say in how their country is governed, which can encourage greater political engagement and help strengthen democratic institutions. They also act as a stable mechanism for the transfer of power between the political parties. Overall, democratic elections are essential for promoting representation and participation as well as ensuring a stable political life.

During an election campaign, political parties compete with each other by generating different policies and ideas to maximize their voting share. Often, political parties decide on their actions and policies with the help of polls. One type of poll which is used by the political parties is the election polls which give us an estimation regarding each party's vote percentage using a random sample of voters representative of the general population. Election polls are critical for political parties as well as the public and other stakeholders. They provide a prediction about the outcome of the election. Parties themselves can use this information to decide on their future actions regarding election campaigns and policy development. For example, a party may choose to change their views on a legislature if they see a decrease in their vote share. Other stakeholders such as businesses and financial institutions often base their investment decisions on the results of an election. Thus, they may rely on predictions to determine their strategy. Finally, one of the most interesting effects of the election polls is on the voters themselves. It is very common for voters to vote for the party which they believe will win the election or receive more votes than they expected. This positive feedback effect is defined as the "Bandwagon effect". Thus, election polls may have a direct influence on the election results as well.

Parties may also choose to use election polls to manipulate public opinion and change the election result. Parties can use election polls that are positively biased towards them to show themselves as winning or their votes as increasing. As a result of the bandwagon effect, indecisive voters decide to vote for the winning or trending party. Therefore, election polls play a critical role in an election process by both predicting and influencing the election results.

The Republic of Türkiye (Turkey) is going to hold its 28th general election on May 14th, 2023. Turkey usually experiences high participation in the elections with voter turnout rates usually higher than 85%. Currently, there are 5 large parties that receive the majority of the votes. The ruling party is the conservative "AKP" (Justice and Development Party) with the support of the nationalist "MHP" (Nationalist Movement Party). The largest party in the opposition is the social democrat "CHP" (Republican People's Party) followed by the center-right "IYI" (Good Party). Another major party in the opposition is the pro-minority left-wing party "HDP" (People's Democratic Party). Turkish election laws allow parties to form alliances for the elections. Currently, AKP and MHP are in an alliance together. The other biggest alliance is formed by CHP and IYI also including the other minor opposition parties.

As the election is expected to be really close, the public gives close attention to the election polls. The number of polls published is relatively high. 55 poll results are published in the first 4 months of 2023. These polls are conducted by various polling companies some of which have questionable credibility. Another issue with election polls in Turkey, it is often speculated that

certain polling companies are positively and/or negatively biased toward certain political parties. Officials from different parties often claim that the results that favor their own party are true and dismiss the results that favor the opposing parties. This makes it very hard for the public, party officials, and other stakeholders to accurately analyze the poll results.

The objective of this project is to understand the bias in the election polls conducted by different polling companies. Depending on the methods used, we will try the estimate the value or the direction (positive or negative) of the bias towards a certain political party from a certain polling company. Whenever the models allow an estimate for the true vote percentage of a political party will also be provided.

2 Data Source

Data for the election polls can be found in Wikipedia and in the publications of the polling companies. Wikipedia is used as the main source for polling data and the official publications are used to verify the data. The study is focused on the largest five parties in Turkey (AKP, CHP, İYİ, MHP, and HDP). These parties received ∼98% of the votes in the last general election in 2018 and in the polls receive 90.3% of the votes on average.

This study covers the polls conducted between January 1st, 2022, and April 21st, 2023. There are 28 different poll companies and 217 polls in the dataset. A preview of the dataset can be seen in Figure 1.

2.1 Date Preprocessing

The first steps in the analysis are exploratory data analysis and data preprocessing. First, the validity of the data is checked by comparing the information in the dataset with the original publication of the polling company. If there are no proper publications found online, the sample is removed from the dataset.

Second, the missing data is completed. Luckily there are no missing instances for the poll results. However, some observations in the data set are missing the date information which is needed by some of the models. This requires further research to determine the date when the survey is conducted. If possible, some assumptions may be used to complete the missing date information. For example, some polling companies publish periodical survey results (e.g. monthly surveys) and conduct their polls at a certain period of each month. If date information is missing for some of their polls then, it can be assumed that these polls are also conducted during a similar time period. Moreover, poll companies usually conduct their polls over a period of time, and we need date information to be a single value, especially for time series analysis methods. Thus, it is assumed that the polls are conducted in the middle of this period. For instance, if a poll is conducted between April 1 and April 5, then it is assumed that the date of the poll is April 3. If there is no information available regarding when the survey is conducted but the date of publication is available, then it is assumed that the polls are conducted 5 days before the publication date. Finally, it is important to note that the approaches used in this analysis do not require date information to be exact, however, approximate dates are used by some of the approaches.

	Company	AKP	CHP	IYI	MHP	HDP	Date	Flag
0	TUSIAR	37.6	26.8	8.7	8.5	9.3	2023-04-21	0
1	Artibir	32.3	32.3	10.2	6.4	11.9	2023-04-20	0
2	Bulgu	34.6	33.9	8.3	5.3	8.7	2023-04-19	0
3	ORC	32.8	28.6	15.1	6.3	9.3	2023-04-20	0
4	Aksoy	31.1	30.6	11.4	7.5	10.5	2023-04-12	1

Figure 1: Data for the five newest poll results

The time plot for the vote rates of the five parties is displayed in Figure 2. As we can see, there is a trend in some of the time series. The linear trend is removed from the series to ensure that the mean of the series is constant during the time of the study. It is assumed that the trend is linear. A time series regression is fitted to each time series and the resulting trend component is extracted from the series.

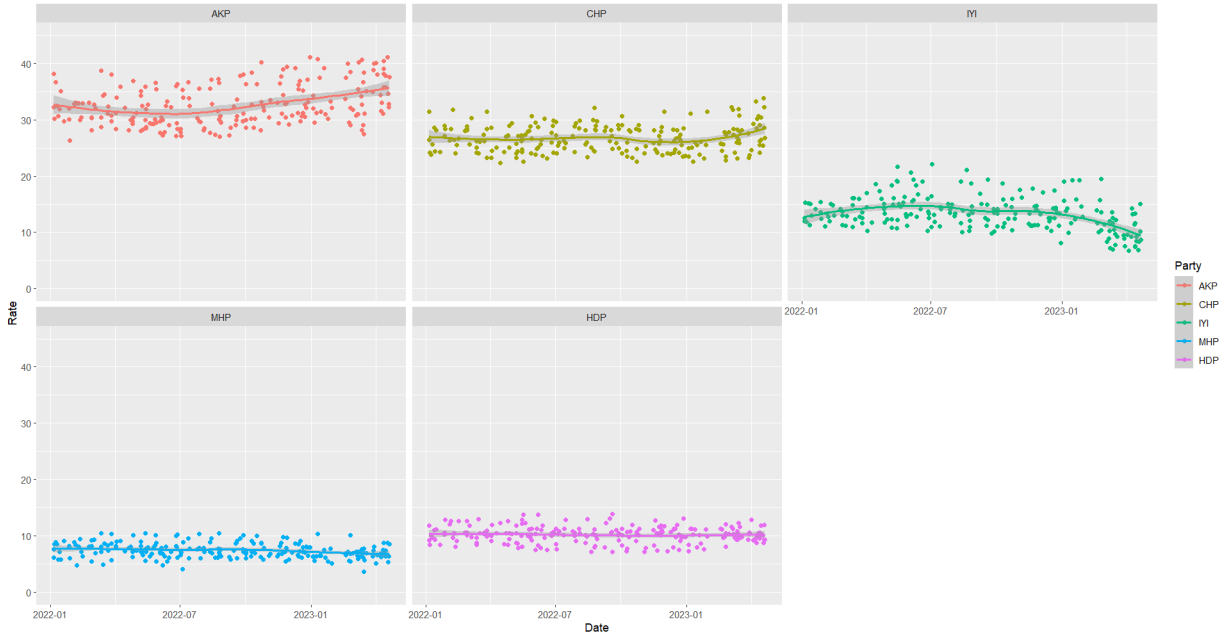

Figure 2: Rate vs. time plot of the poll results for each political party

Normality of the data for each political party is also important as for most of the models we need data to follow the normal distribution. Histograms of the vote percentages of each party are displayed in Figure 3. For some of the parties, the data is far from the normal distribution. Thus, data is standardized during preprocessing. Time plots and histograms for the detrended and standardized time series can be seen in Figures 4 and 5. After preprocessing, data is stationary as there is no trend or seasonality in the data. Also, the data is approximately normal. However, the histogram for the largest two parties slightly deviates from normal distribution as they are right-skewed. These deviations are relatively small and do not affect the output of the models. Thus, we can continue with the implementation of different models.

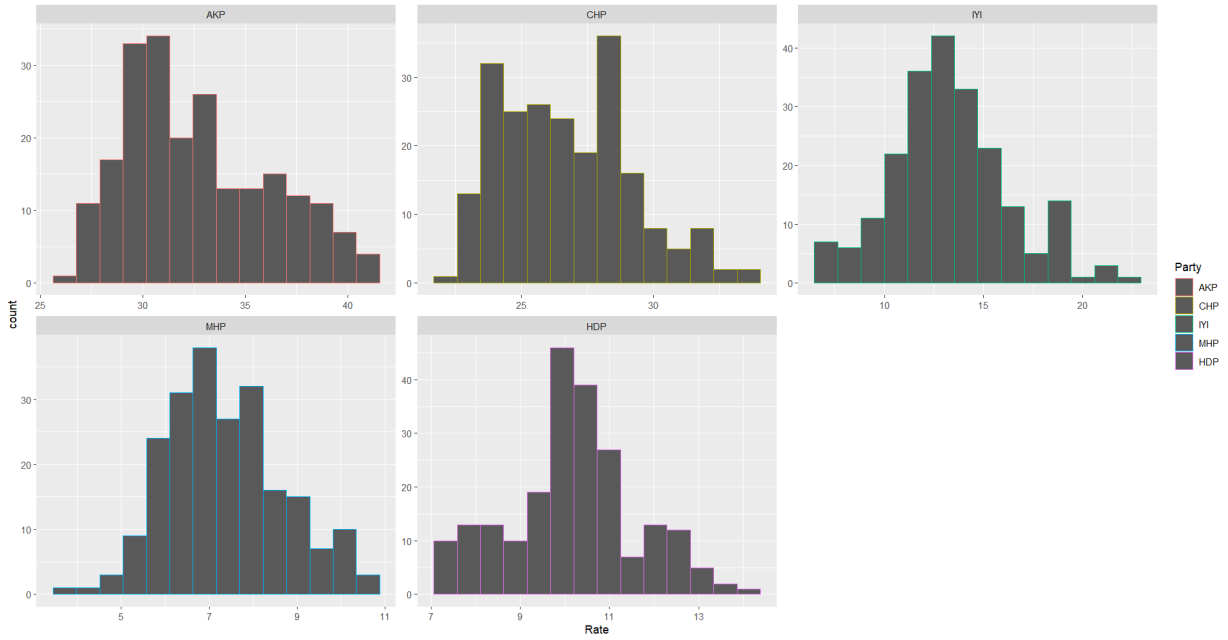

Figure 3: Histogram of the poll results for each political party

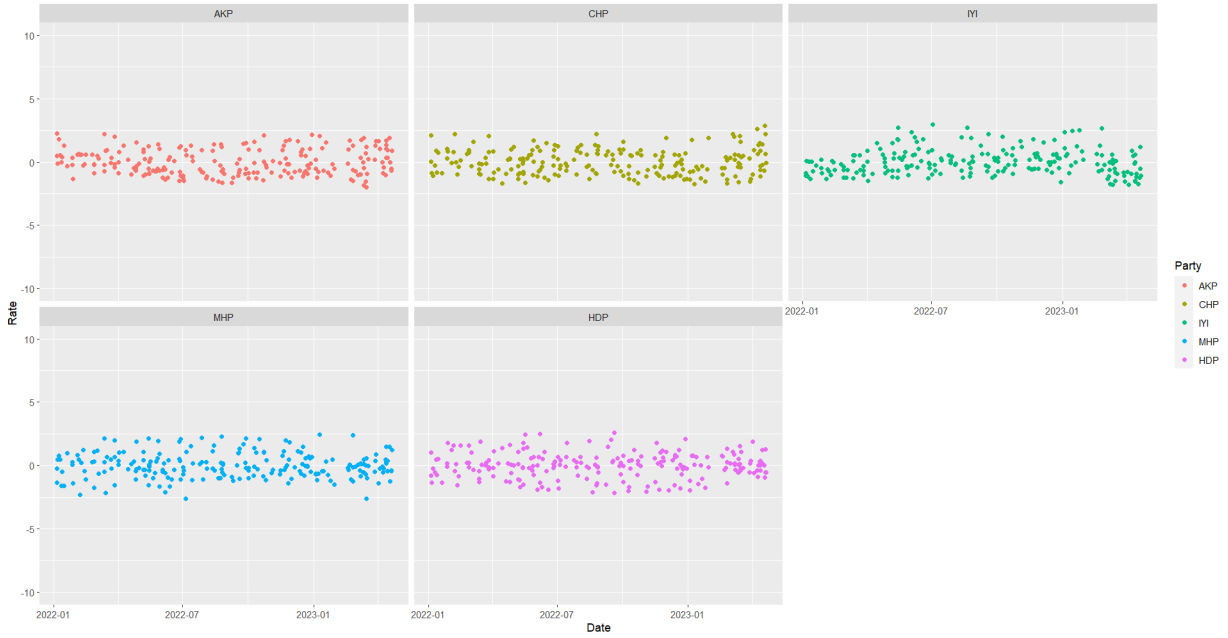

Figure 4: Detrended and standardized poll results vs. time plot of the poll results for each political party

3 Methodology

After exploratory data analysis and preprocessing to further understand the characteristics of the data and ensure stationarity and normality, different approaches are used to analyze the bias of the polling companies.

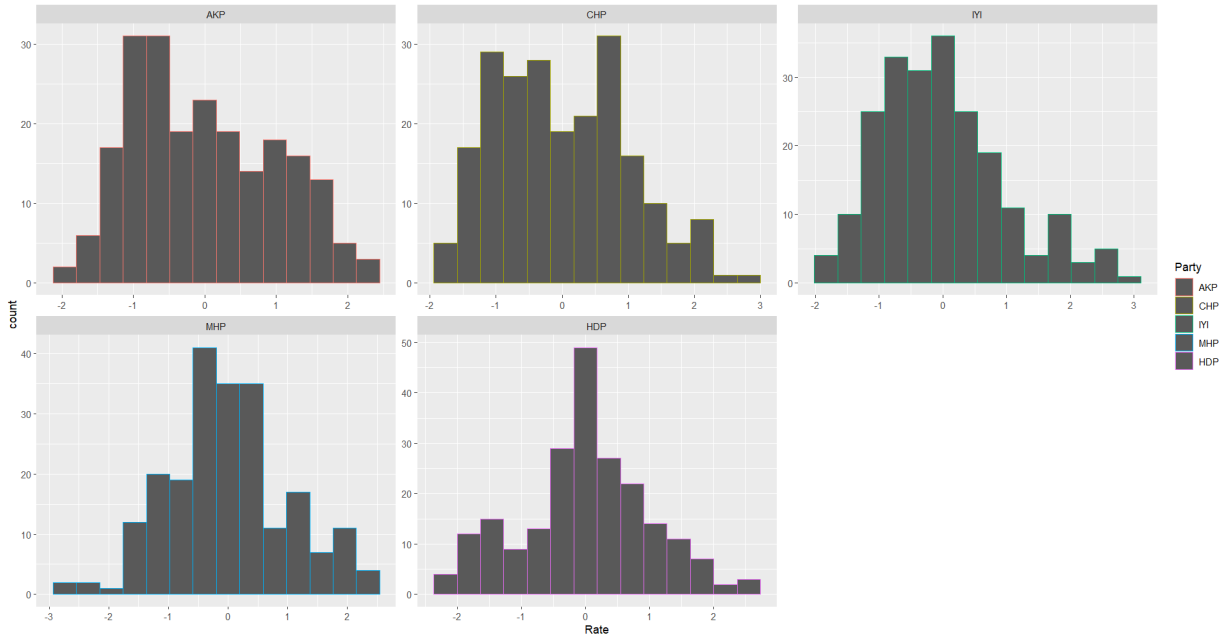

Figure 5: Histogram of the detrended and standardized poll results for each political party

In order to understand the bias, several different machine-learning methods are used from the Computational Statistics class. Other statistical methods such as time series regression are also used to understand the trend, and bias in the data. The first method that will be used to understand the bias is the **EM Algorithm**.

Second, clustering algorithms are used to estimate the direction of the bias between the companies and political parties. We use clustering algorithms such as **K-Means** and **Gaussian Mixture Models** to cluster the observations (polls) into different clusters. Here, each cluster represents a different behavior for the poll companies. For example, for a cluster, we can deduce that the polls in this cluster are positively biased for parties X, Y, etc., and negatively biased for parties A, B, etc.

Finally, we evaluate the data as time series and use different time series models. Models such as **Smoothing Splines** from Computational Statistics. Then, using interpolation we can estimate the bias of a certain company towards a political party. Detailed methodology for each of the approaches is explained in the following sections. Evaluation of the results can be found in the final section.

3.1 Expectation-Maximization

An EM-Algorithm which is similar to the EM application developed by John Platt is implemented. In this case, instead of papers and reviewers, we have political parties and different poll companies. In this setting, each party will have a normally distributed true "rating". Furthermore, each polling company will have a normally distributed bias towards a political party. Then, each realization (an election poll) can be expressed as the sum of these two effects

and a random error. In this model, it is assumed that a polling company's (reviewer) bias towards different parties (papers) has the same distribution ($N(\mu_p, \sigma_p^2)$) so we need to have the same bias distribution between a polling company and the government and opposition parties. This would not be realistic since it is often speculated that a polling company often has a positive bias towards one side and a negative bias towards the other. Then, it is more realistic to assume that there are two different models. The first model is between the polling companies and the government parties and the second model is between the polling companies and the opposition parties. We assume that the bias of a company towards the government and opposition parties are independent of each other. In this model, we also assume that the rating given by a polling company p to a party r, $x_{(pr)}$, is generated with the following process.

$$\begin{aligned} z_p &\sim N(\mu_p, \sigma_p^2) \\ y_r &\sim N(\nu_r, \tau_r^2) \\ x_{(pr)}|z_{(pr)}, y_{(pr)} &\sim N(y_{(pr)} + z_{(pr)}, \sigma^2) \end{aligned}$$

Here, $x_{(pr)}$ is the rating given by company p to party r, $z_{(pr)}$ is the realization of the bias of company p to r and $y_{(pr)}$ is the realization of the true rating of the party r. We can represent $x_{(pr)}$ as

$$x_{(pr)} = z_{(pr)} + y_{(pr)} + \epsilon_{(pr)}$$

where $\epsilon_{(pr)}$ is the independent random noise with normal distribution. With this process, we can see that the joint distribution of $x_{(pr)}, z_{(pr)}, y_{(pr)}$ is multivariate normal.

$$x_{(pr)}, z_{(pr)}, y_{(pr)} \sim \mathbf{N}(\begin{bmatrix} \mu_p \\ \nu_r \\ \mu_p + \nu_r \end{bmatrix}, \begin{bmatrix} \sigma_p^2 & 0 & \sigma_p^2 \\ 0 & \tau_r^2 & \tau_r^2 \\ \sigma_p^2 & \tau_r^2 & \sigma_p^2 + \tau_r^2 + \sigma^2 \end{bmatrix})$$

Then, the conditional distribution, $z_{(pr)}, y_{(pr)}|x_{(pr)}$ is also normal and can be expressed as

$$Q_{pr} = p(z_{(pr)}, y_{(pr)}|x_{(pr)}) = \mathbf{N}(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix})$$

and the parameters of the conditional distribution are equal to

$$\begin{aligned} \mu_1 &= \mu_p + \frac{\sigma_p^2}{\sigma_p^2 + \sigma^2 + \tau_r^2}(x_{(pr)} - \mu_p - \nu_r) \\ \mu_2 &= \nu_r + \frac{\tau_r^2}{\sigma_p^2 + \sigma^2 + \tau_r^2}(x_{(pr)} - \mu_p - \nu_r) \\ \Sigma' &= \begin{bmatrix} \sigma_p^2 & 0 \\ 0 & \sigma_p^2 \end{bmatrix} - \begin{bmatrix} \sigma_p^2 \\ \tau_r^2 \end{bmatrix} \frac{1}{\sigma_p^2 + \sigma^2 + \tau_r^2}(x_{(pr)} \begin{bmatrix} \sigma_p^2 & \tau_r^2 \end{bmatrix}) \end{aligned}$$

E-step of the EM algorithm consists of the computation of the parameters of the conditional distribution. Then, in the M-step, parameters for the distribution of the latent variables are calculated using the following update rules. Detailed derivations of the conditional distribution

and the update rules can be found in the appendix.

$$\mu_p' = \frac{1}{R}\sum_{i=1}^{R}\mu_1$$
$$\nu_r' = \frac{1}{P}\sum_{i=1}^{P}\mu_2$$
$$(\sigma_p')^2 = \frac{1}{R}\sum_{i=1}^{R}\left(\Sigma_{11} + \mu_1^2 - 2\mu_1\mu_p + \mu_p^2\right)$$
$$(\tau_r')^2 = \left(\Sigma_{22} + \mu_2^2 - 2\mu_2\nu_r + \nu_r^2\right)$$

For this approach, only the most recent survey from each company is used. The data is standardized before being passed to the model. EM algorithm requires initial points and a stopping condition to be selected as hyperparameters. The algorithm stops when the change in the likelihood between successive iterations is less than 0.0001. The initial points of the two algorithms are determined by trying out different starting points. As a result, the algorithm converged to a locally optimal solution. The results of this algorithm are evaluated in the next section.

There are possible extensions to this model. First, the algorithm only takes one survey from each company into account. The model can be extended to include multiple surveys from each company. Also, we assume that the biases towards the opposition and government parties are formed independently which is not realistic. Instead, we can assume that there is dependence between the biases and develop an EM method using multivariate normal distributions for the variables $x_{(pr)}, z_{(pr)}, y_{(pr)}$.

3.2 Gaussian Mixture Models

The second approach is to use clustering algorithms to determine different behaviors between the polling companies. The advantage of using a clustering algorithm is that we are able to use data from each poll together instead of constructing different models for each party as we did in the EM application and cubic splines. This allows us to take into account the dependencies between different parties.

The first clustering algorithm used is the Gaussian Mixture Models. Data is detrended and standardized in the preprocessing steps. All available data is included in the model so there may be more than one poll from each company. There were several alternatives for the number of Gaussian components. Models with 2, 3, and 5 components are considered. Each of these models had a different real-world interpretation. Two-component model is likely to divide the polling companies between the government and the opposition. Three-component model is likely to divide the companies between the three major alliances. Finally, using 5 components is reasonable since there are 5 different parties in the model. Different models are trained to select the most plausible interpretation. As a result, the model with 5 components is selected.

After clustering the data, each poll is assigned to the clusters with corresponding weights. It is possible that polls from the same company are assigned to different clusters. Thus, assignment weights are averaged over all the polls of a company and it is assigned to the cluster with the

highest average weight. Each cluster has a mean vector whose elements correspond to the 5 parties. Using the mean vectors, we can deduce the direction of the bias. For example, for a cluster, the mean corresponding to a party is positive means that the companies assigned to that cluster have a positive bias towards that party. Note that since the data is standardized, the sign of the mean vector is the same as the direction of the bias. The results of this approach are evaluated in the next section.

3.3 K-Means Clustering

The second clustering method used is the K-Means. Similar to the GMM, data is detrended and standardized. Also, the data is divided into 5 different clusters. Since the K-Means algorithm does not provide assignment weights and makes a hard assignment between the samples and clusters, we need to use a different method to determine the cluster of a company. Majority voting is used to determine the cluster of a company. A company is assigned to a cluster if the majority of its polls are assigned to that cluster. The directions of bias for each cluster can be deduced by using the signs of the cluster means. If there is a tie in the majority voting, then the mean vectors of the tied clusters are averaged to compute the direction of the bias. The results of this approach are evaluated in the next section.

3.4 Cubic Splines

The final approach used to model the bias is from the domain of time series regression. In the previous approaches, the time of the survey is not considered. However, it can be argued that valuable insights regarding the data are lost by not taking into account the time aspect. Thus, using a model that takes the date information into account can provide more valuable insights.

Cubic splines are used to analyze the time series. We have 5 different time series corresponding to the 5 parties. Bias between a company and a party is considered independent of the other biases. The methodology used in this approach is as follows: For a party-company pair:

1. Detrend the time series (i.e. remove the linear trend)
2. Remove the polls of the company from the time series
3. Compute the monthly averages
4. Fit a cubic spline model to the monthly averages
5. Calculate bias as Actual Value - Predicted Value for each poll of the company
6. Average the bias of each poll to determine the overall bias and its direction

Monthly averages are considered instead of the actual values because the data is very noisy as you can see in Figure 4 and it becomes very hard to fit a cubic spline that represents the trends in the data. By averaging the multiple entries we can smooth the time series. The number of knots is chosen as 4. Each interval corresponds to a quarter (4 from 2022 and 1 from 2023). The value of the penalty term is determined using cross-validation. Here, we demonstrate how the approach works using an example. We select the company named "MetroPOLL" and the party "AKP". We first remove the polls of "MetroPOLL" from the series and take the monthly averages (Figure 6, black points). Then, we fit a cubic spline to the data (Figure 6, black line). Finally, we calculate the difference between the red dots and the black line in Figure 6. As you can see, red points lie

above the fitted line which suggests that the pollster "MetroPOLL" is positively biased towards the party "AKP".

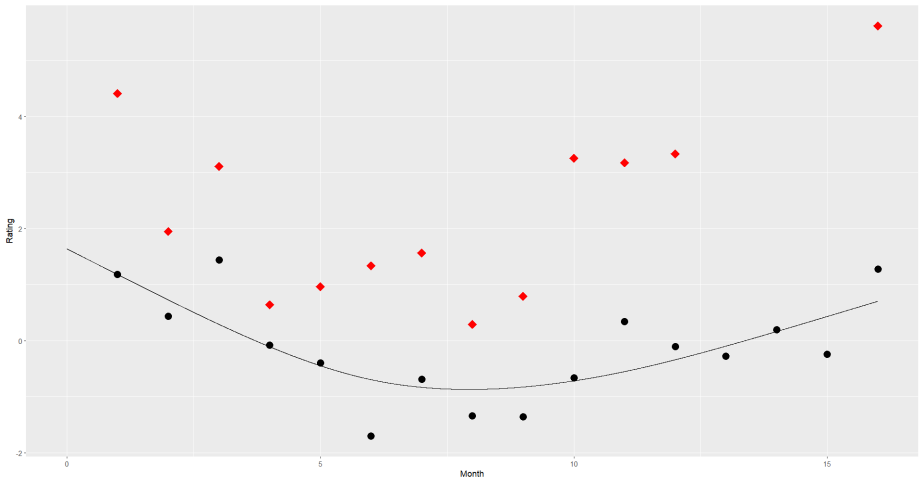

Figure 6: Cubic Spline fit for "MetroPOLL"-"AKP"

4 Evaluation and Final Results

Since we do not know the true values for the biases, it is not easy to measure the performance of an algorithm. Thus, one way to comment on the performance of the models is to use the prior public speculation about the companies and compare them with the results of the algorithm. There are some companies which are known to be highly biased. For example, "GENAR" is known to be biased towards the "AKP", whereas the "Yoneylem" is known to have a negative bias towards the "AKP". Using such cases, we tried to understand the performance of the algorithm. Overall, the model results are parallel to the prior predictions the public had regarding these companies. Detailed results for each algorithm can be found in the corresponding section in the Appendix. Here we provide the key insights from each algorithm.

4.1 Expectation-Maximization

Expectation-Maximization algorithms both for the government and opposition parties converged to locally optimal solutions. Plots of the expected likelihood vs iteration can be seen in Figure 7 and 8.
The true mean of the parties, ν_r converged to the following values (Table 1). These results are really close to the mean survey rating for each political party which suggests that there is an equal number of positively and negatively biased parties towards the opposition and the government.

	Government		Opposition		
	AKP	MHP	CHP	IYI	HDP
Mean	31.25	6.37	25.09	10.80	9.64

Table 1: Final means for each party

For the government parties, the companies with the highest positive bias are "Global & Akademetre" and "GENAR" and the companies "Themis" and "Yoneylem" have the highest

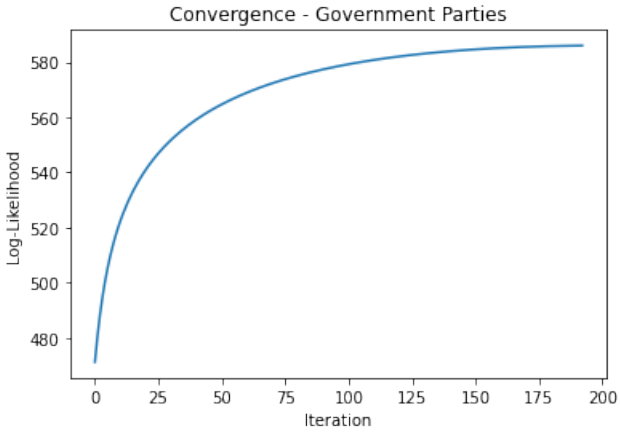

Figure 7: Convergence of the model for the government parties

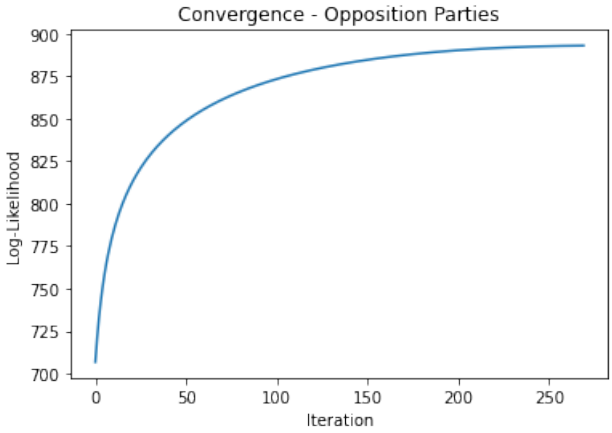

Figure 8: Convergence of the model for the opposition parties

negative bias. For the opposition parties, the companies with the highest positive bias are "Sosyo-Politik" and "KONDA" whereas the companies with the lowest negative bias are "Areda Survey" and "ORC". Overall, the company with the lowest total bias (sum of the absolute values of the biases") are "IEA", "SAROS" and "Area". On the other hand, the companies with the highest total bias are "Global & Akademetre", "GENAR" and "Themis". Note that the bias values presented are for the standardized data, therefore do not represent the true value of the bias. However, we can still make inferences regarding the relative bias.

The distribution of bias between the companies and parties can be seen in Table 2. Interestingly, the number of positively biased companies are less than the number of negatively biased companies and there are 9 companies with negative bias towards both sides. These results suggest that for some companies (positive-positive and negative-negative bias), the bias does not depend on political alignment but it depends on the method of the survey.

Count		Opposition	
		Positive	Negative
Government	Positive	4	7
	Negative	8	9

Table 2: Distribution of Bias

4.2 Gaussian Mixture Models

After the clustering using GMM, mean vectors of the components are used to determine the bias. By looking at the sign of the mean vector for a particular party, we can determine the direction of the bias. The resulting mean vectors for each component and the number of assigned companies to each component can be seen in Table 3. Components 1 and 5 represent the companies that are positively biased towards the main opposition party "CHP". On the other hand, components 2 and 3 represents the companies that are positively biased towards the government, "AKP". Component 4 consists of the companies that favor the second opposition party, "IYI".

	AKP	CHP	IYI	MHP	HDP	Number of Companies
Component 1	-0.56	0.56	0.08	-0.01	-0.13	9
Component 2	0.21	-0.20	-0.05	-0.40	0.97	9
Component 3	1.36	-0.80	-0.99	1.27	0.18	7
Component 4	-0.95	-0.44	1.58	-0.59	-1.60	2
Component 5	-0.47	1.81	-0.49	-1.19	-0.58	1

Table 3: Mean Vectors for the GMM

In Table 4, we can see how the biased companies are distributed between parties. As expected, the main government party "AKP" has more positively biased companies. On the other hand, the main opposition party "CHP" has more negatively biased companies. It is known that "AKP" uses poll results as a tactic the influence public opinion.

Parties	Positive Bias	Negative Bias
AKP	16	12
CHP	10	18
IYI	11	17
MHP	7	21
HDP	16	12

Table 4: Distribution of Bias among the Parties

There are multiple companies assigned to each component so we cannot determine the companies with the highest and lowest biases. However, by looking at the mean vectors, we can deduce that the companies assigned to component 1 and component 2 are less biased compared to the companies assigned to the other components since the sum of absolute values for these components is smaller.

4.3 K-Means Clustering

K-Means clustering provides similar outputs to the GMM. By looking at the sign of the cluster means we can determine the bias directions. The resulting cluster means and component and the number of assigned companies to each vector can be seen in Table 5. Clusters of the K-Means have a similar interpretation to the GMM. Two clusters consist of companies that favor "AKP" (clusters 3 and 4) and two other clusters favor "CHP", (clusters 2 and 5). On the other hand, one cluster (Cluster 1) stands out as it favors the "IYI" by a large margin.

Cluster	AKP	CHP	IYI	MHP	HDP	Number of Companies
1	-0.91	-0.64	1.70	-0.56	-1.51	3
2	-0.56	1.61	-0.44	-0.76	-0.03	3
3	0.05	-0.20	0.10	-0.75	1.38	6
4	1.34	-0.80	-0.91	1.14	0.25	7
5	-0.43	0.24	0.11	0.16	-0.23	5
Tie	-	-	-	-	-	4

Table 5: Cluster Means

Contrary to the GMM, the two biggest parties, "AKP" and "CHP", have more negatively biased companies than positively biased companies. One notable difference is in the biases towards the two nationalist parties "IYI" and "MHP". K-means result in more positively biased companies for both of these parties.

Parties	Positive Bias	Negative Bias
AKP	13	15
CHP	9	19
IYI	17	11
MHP	14	14
HDP	13	15

Table 6: Distribution of Bias among the Parties

4.4 Cubic Splines

The approach we used with the cubic splines allows us to determine the companies with the highest and lowest biases can be determined. In Table 7, the top three companies with the highest and lowest biases for each party are shown. Parties with a bias towards a government or opposition party often have the opposite bias towards the other side. For example, "Avrasya" has the highest positive bias towards "CHP", whereas it has the highest negative bias towards "AKP".

Party	Positive Bias			Negative Bias		
AKP	ADA	Areda Survey	Global & Akad.	ALF	ORC	Avrasya
CHP	Avrasya	Objektif	Yoneylem	Sonar	GENAR	ADA
IYI	KONDA	ORC	REMRES	ADA	Optimar	GENAR
MHP	Areda Survey	Global & Akad.	Optimar	Themis	REMRES	Sosyo-Politik
HDP	Sosyo-Politik	MetroPOLL	SAROS	ORC	MAK	SONAR

Table 7: Companies with Highest and Lowest Bias

Finally, the distribution of the bias among the parties is as follows:

Parties	Positive Bias	Negative Bias
AKP	13	15
CHP	14	14
IYI	12	16
MHP	12	16
HDP	17	11

Table 8: Distribution of Bias among the Parties

When we look at the total bias a company has, we see that "ADA", "Areda Survey" and "KONDA" has the highest bias. On the other hand, "IEA", "Bulgu" and "Area" has the lowest total bias.

4.5 Conclusion

All in all, it is not possible to measure exactly the accuracy of the models since the election will be held on the 14th of May. On the other hand, "sanity" checks using the existing public opinion can be performed to understand the performance of the algorithm. Once we have the actual election results in May, we will be able to comment on the performance of the algorithms.

However, the EM algorithm and the cubic splines produced similar results for the least biased companies. Based on these algorithms, "IEA" and "Area" are the least biased companies.

5 References

- https://www.lse.ac.uk/research/research-for-the-world/impact/the-politics-of-polling-why-are-polls-important-during-elections
- https://tr.wikipedia.org/wiki/%C3%9Clke_%C3%A7ap%C4%B1nda_2023_T%C3%BCrkiye_genel_se%C3%A7imleri_i%C3%A7in_yap%C4%B1lan_anketler

6 Appendix

6.1 Expectation Maximization Results

Company	Bias - Government	Bias - Opposition	Total
ADA	0.00561	-0.00056	0.006175
Aksoy	-0.00041	0.00010	0.000519
ALF	-0.00444	-0.00054	0.004982
Area	-0.00009	0.00026	0.000345
Areda Survey	0.00643	-0.00062	0.007051
Artibir	-0.00374	0.00042	0.004163
ASAL	0.00290	-0.00030	0.003202
Avrasya	-0.00592	-0.00001	0.005929
Bulgu	-0.00346	-0.00021	0.003671
GENAR	0.00726	-0.00034	0.007595
Gezici	0.00425	0.00031	0.004554
Global & Akademetre	0.01222	0.00043	0.012650
IEA	-0.00007	0.00003	0.000104
KONDA	0.00049	0.00119	0.001682
MAK	-0.00250	-0.00037	0.002872
MetroPOLL	-0.00233	-0.00015	0.002481
Objektif	-0.00201	0.00086	0.002864
Optimar	0.00365	-0.00054	0.004195
ORC	-0.00477	-0.00063	0.005406
Piar	-0.00250	-0.00037	0.002871
Remres	-0.00238	0.00092	0.003303
SAROS	-0.00009	0.00018	0.000266
SONAR	0.00217	-0.00044	0.002609
Sosyo Politik	-0.00397	0.00121	0.005184
TEAM	0.00297	0.00009	0.003057
Themis	-0.00809	-0.00035	0.008446
TUSIAR	0.00492	-0.00040	0.005323
Yoneylem	-0.00597	-0.00016	0.006129

Table 9: Estimated Biases from the EM Algorithm

6.2 Gaussian Mixture Models Results

Company	Cluster	AKP	CHP	IYI	MHP	HDP	Sample Size
ADA	3	1	-1	-1	1	1	3
ALF	4	-1	-1	1	-1	-1	15
ASAL	3	1	-1	-1	1	1	13
Aksoy	1	-1	1	1	-1	-1	16
Area	1	-1	1	1	-1	-1	11
Areda Survey	3	1	-1	-1	1	1	13
Artibir	2	1	-1	-1	-1	1	8
Avrasya	1	-1	1	1	-1	-1	14
Bulgu	1	-1	1	1	-1	-1	5
GENAR	3	1	-1	-1	1	1	2
Gezici	1	-1	1	1	-1	-1	2
Global & Akademetre	3	1	-1	-1	1	1	1
IEA	2	1	-1	-1	-1	1	4
KONDA	2	1	-1	-1	-1	1	1
MAK	1	-1	1	1	-1	-1	11
MetroPOLL	2	1	-1	-1	-1	1	13
ORC	4	-1	-1	1	-1	-1	22
Objektif	5	-1	1	-1	-1	-1	1
Optimar	3	1	-1	-1	1	1	11
Piar	1	-1	1	1	-1	-1	12
Remres	1	-1	1	1	-1	-1	1
SAROS	2	1	-1	-1	-1	1	7
SONAR	2	1	-1	-1	-1	1	3
Sosyo Politik	2	1	-1	-1	-1	1	2
TEAM	2	1	-1	-1	-1	1	4
TUSIAR	3	1	-1	-1	1	1	3
Themis	2	1	-1	-1	-1	1	3
Yoneylem	1	-1	1	1	-1	-1	16

Table 10: Estimated Bias Directions from the GMM

6.3 K-Means Clustering Results

Company	Cluster	AKP	CHP	IYI	MHP	HDP	Sample Sizes
ADA	4	1	-1	-1	1	1	3
ALF	1	-1	-1	1	-1	-1	15
ASAL	4	1	-1	-1	1	1	13
Aksoy	5	-1	1	1	1	-1	16
Area	5	-1	1	1	1	-1	11
Areda Survey	4	1	-1	-1	1	1	13
Artibir	3	1	-1	1	-1	1	8
Avrasya	2	-1	1	-1	-1	-1	14
Bulgu	5	-1	1	1	1	-1	5
GENAR	4	1	-1	-1	1	1	2
Gezici	Tie	-1	-1	-1	1	-1	2
Global & Akademetre	4	1	-1	-1	1	1	1
IEA	Tie	-1	-1	1	-1	-1	4
KONDA	3	1	-1	1	-1	1	1
MAK	5	-1	1	1	1	-1	11
MetroPOLL	3	1	-1	1	-1	1	13
ORC	1	-1	-1	1	-1	-1	22
Objektif	2	-1	1	-1	-1	-1	1
Optimar	4	1	-1	-1	1	1	11
Piar	5	-1	1	1	1	-1	12
Remres	1	-1	-1	1	-1	-1	1
SAROS	3	1	-1	1	-1	1	7
SONAR	Tie	-1	-1	1	1	-1	3
Sosyo Politik	3	1	-1	1	-1	1	2
TEAM	3	1	-1	1	-1	1	4
TUSIAR	4	1	-1	-1	1	1	3
Themis	Tie	-1	1	1	-1	-1	3
Yoneylem	2	-1	1	-1	-1	-1	16

Table 11: Estimated Bias Directions from the K-Means

6.4 Cubic Spline Results

Company	AKP	CHP	IYI	MHP	HDP	Sum of the Absolute Values
ADA	6.32	-3.14	-3.99	0.59	0.03	14.06
Aksoy	-2.06	1.96	0.01	0.37	0.32	4.72
ALF	-4.07	1.11	2.41	-0.80	-1.42	9.82
Area	-0.68	0.49	1.33	0.38	0.34	3.21
Areda Survey	5.83	-2.84	-2.85	2.67	0.28	14.47
Artibir	-3.03	1.90	0.54	-1.32	1.80	8.59
ASAL	3.81	-1.64	-2.14	1.13	0.29	9.01
Avrasya	-3.24	4.24	-1.00	-1.38	0.46	10.32
Bulgu	-0.99	2.03	-0.48	-0.84	-0.25	4.57
GENAR	4.96	-3.19	-3.77	0.67	0.66	13.24
Gezici	1.30	0.94	-2.85	1.37	0.32	6.79
Global & Akademetre	5.82	1.18	-2.67	2.45	0.37	12.48
IEA	0.58	-0.01	0.62	0.27	0.68	2.16
KONDA	-2.71	-1.66	7.35	-0.21	1.93	13.86
MAK	-1.13	-0.55	1.97	-0.60	-1.78	6.04
MetroPOLL	2.47	-2.40	0.05	-0.62	2.33	7.88
Objektif	-1.87	3.68	-0.24	-1.46	-0.66	7.91
Optimar	5.59	-0.90	-3.86	1.60	-0.23	12.19
ORC	-3.50	-2.11	4.87	-0.47	-2.48	13.44
Piar	-1.57	0.43	-2.40	0.75	1.03	6.17
Remres	-1.39	0.87	4.82	-2.14	-0.29	9.50
SAROS	1.15	-0.50	-1.58	-0.50	2.05	5.77
SONAR	1.66	-3.29	2.84	-0.27	-1.48	9.53
Sosyo Politik	-0.91	2.44	0.51	-1.83	3.30	8.99
TEAM	3.85	-0.37	-1.15	-0.03	0.91	6.30
Themis	-1.94	1.58	-1.29	-2.48	-0.86	8.15
TUSIAR	3.02	-2.06	-2.37	0.64	-0.35	8.43
Yoneylem	-0.71	2.44	-0.77	-0.65	-0.46	5.03

Table 12: Estimated Biases from the Cubic Splines

6.5 Derivation of EM Algorithm

Parameters of the conditional distribution:

$$\begin{aligned}
\bar{\mu} &= \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \\
&= \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x^{(pr)} - \mu_2) \\
&= \begin{bmatrix} \mu_p \\ \nu_r \end{bmatrix} + \begin{bmatrix} \sigma_p^2 \\ \tau_r^2 \end{bmatrix} (\sigma_p^2 + \tau_r^2 + \sigma^2)^{-1}(x^{(pr)} - \mu_p - \nu_r) \\
&= \begin{bmatrix} \mu_p + \frac{(x^{(pr)} - \mu_p - \nu_r)}{(\sigma_p^2 + \tau_r^2 + \sigma^2)}\sigma_p^2 \\ \nu_r + \frac{(x^{(pr)} - \mu_p - \nu_r)}{(\sigma_p^2 + \tau_r^2 + \sigma^2)}\tau_r^2 \end{bmatrix}
\end{aligned}$$

$$
\begin{aligned}
\mathbf{\Sigma} &= \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \\
&= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \\
&= \begin{bmatrix} \sigma_p^2 & 0 \\ 0 & \tau_r^2 \end{bmatrix} - \begin{bmatrix} \sigma_p^2 \\ \tau_r^2 \end{bmatrix} (\sigma_p^2 + \tau_r^2 + \sigma^2)^{-1} \begin{bmatrix} \sigma_p^2 & \tau_r^2 \end{bmatrix} \\
&= \begin{bmatrix} \sigma_p^2 & 0 \\ 0 & \tau_r^2 \end{bmatrix} - (\sigma_p^2 + \tau_r^2 + \sigma^2)^{-1} \begin{bmatrix} \sigma_p^4 & \tau_r^2\sigma_p^2 \\ \tau_r^2\sigma_p^2 & \tau_r^4 \end{bmatrix} \\
&= (\sigma_p^2 + \tau_r^2 + \sigma^2)^{-1} \begin{bmatrix} \sigma_p^2(\tau_r^2 + \sigma^2) & -\tau_r^2\sigma_p^2 \\ -\tau_r^2\sigma_p^2 & \tau_r^2(\sigma_p^2 + \sigma^2) \end{bmatrix}
\end{aligned}
$$

M-step updates:

$$
\begin{aligned}
Q_{pr}(\theta'|\theta) &= \mathbb{E}_Q \left[\log p(y^{(p)}, z^{(r)}, x^{(pr)}) | x^{(pr)};\ \theta \right] \\
&= \mathbb{E}_Q \left[\log \left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{(\frac{-1}{2} \frac{(x^{(pr)} - y^{(p)} - z^{(r)})^2}{\sigma^2})} \frac{1}{\sqrt{2\pi\sigma_p^2}} e^{(\frac{-1}{2} \frac{(y^{(p)} - \mu_p)^2}{\sigma_p^2})} \frac{1}{\sqrt{2\pi\tau^2}} e^{(\frac{-1}{2} \frac{(z^{(r)} - \nu_r)^2}{\tau^2})} \right) \right] \\
&= \mathbb{E}_Q \left[\log \left(\frac{1}{(2\pi)^{3/2}\sigma\sigma_p\tau_r} e^{(\frac{-1}{2} \frac{(x^{(pr)} - y^{(p)} - z^{(r)})^2}{\sigma^2})} e^{(\frac{-1}{2} \frac{(y^{(p)} - \mu_p)^2}{\sigma_p^2})} e^{(\frac{-1}{2} \frac{(z^{(r)} - \nu_r)^2}{\tau_r^2})} \right) \right] \\
&= \mathbb{E}_Q \big[\log \left(\frac{1}{(2\pi)^{3/2}\sigma\sigma_p\tau_r} \right) \\
&- (\frac{(x^{(pr)} - y^{(p)} - z^{(r)})^2}{2\sigma^2}) - (\frac{((y^{(p)})^2 - 2y^{(p)}\mu_p + \mu_p^2)}{2\sigma_p^2}) - (\frac{(z^{(r)})^2 - 2z^{(r)}\nu_r + \nu_r^2}{2\tau_r^2}) \big] \\
&= \mathbb{E}_Q \left[\log \left(\frac{1}{(2\pi)^{3/2}\sigma\sigma_p\tau_r} \right) - (\frac{(x^{(pr)} - y^{(p)} - z^{(r)})^2}{2\sigma^2}) \right] \\
&- \mathbb{E}_Q \left[(\frac{((y^{(p)})^2 - 2y^{(p)}\mu_p + \mu_p^2)}{2\sigma_p^2}) \right] - \mathbb{E}_Q \left[(\frac{(z^{(r)})^2 - 2z^{(r)}\nu_r + \nu_r^2}{2\tau_r^2}) \right] \\
&= \mathbb{E}_Q \left[\log \left(\frac{1}{(2\pi)^{3/2}\sigma\sigma_p\tau_r} \right) - \left(\frac{(x^{(pr)} - y^{(p)} - z^{(r)})^2}{2\sigma^2} \right) \right] \\
&- (\frac{\mathbb{E}_Q \left[(y^{(p)})^2 \right] - 2\mathbb{E}_Q \left[y^{(p)} \right] \mu_p + \mu_p^2}{2\sigma_p^2}) - (\frac{\mathbb{E}_Q \left[(z^{(r)})^2 \right] - 2\mathbb{E}_Q \left[z^{(r)} \right] \nu_r + \nu_r^2}{2\tau_r^2}) \\
&= \mathbb{E}_Q \left[\log \frac{1}{\sigma_p\tau_r} \right] + \mathbb{E}_Q \left[\log \left(\frac{1}{(2\pi)^{3/2}\sigma} \right) - \left(\frac{(x^{(pr)} - y^{(p)} - z^{(r)})^2}{2\sigma^2} \right) \right] \\
&- (\frac{\mathbb{E}_Q \left[(y^{(p)})^2 \right] - 2\mathbb{E}_Q \left[y^{(p)} \right] \mu_p + \mu_p^2}{2\sigma_p^2}) - (\frac{\mathbb{E}_Q \left[(z^{(r)})^2 \right] - 2\mathbb{E}_Q \left[z^{(r)} \right] \nu_r + \nu_r^2}{2\tau_r^2})
\end{aligned}
$$

$$
\begin{aligned}
\theta' &= \underset{\theta'}{\operatorname{argmin}} \sum_{r=1}^{R}\sum_{p=1}^{P} \mathbb{E}_Q\left[\log p(y^{(p)}, z^{(r)}, x^{(pr)}) | x^{(pr)};\ \theta\right] \\
&= \underset{\theta'}{\operatorname{argmin}} \sum_{r=1}^{R}\sum_{p=1}^{P} \mathbb{E}_Q\left[\log \frac{1}{\sigma_p \tau_r}\right] + \mathbb{E}_Q\left[\log\left(\frac{1}{(2\pi)^{3/2}\sigma}\right) - \left(\frac{(x^{(pr)} - y^{(p)} - z^{(r)})^2}{2\sigma^2}\right)\right] \\
&- \left(\frac{\mathbb{E}_Q\left[(y^{(p)})^2\right] - 2\mathbb{E}_Q\left[y^{(p)}\right]\mu_p + \mu_p^2}{2\sigma_p^2}\right) - \left(\frac{\mathbb{E}_Q\left[(z^{(r)})^2\right] - 2\mathbb{E}_Q\left[z^{(r)}\right]\nu_r + \nu_r^2}{2\tau_r^2}\right) \\
&= \underset{\theta'}{\operatorname{argmin}} \sum_{r=1}^{R}\sum_{p=1}^{P} \mathbb{E}_Q\left[\log \frac{1}{\sigma_p \tau_r}\right] - \left(\frac{\mathbb{E}_Q\left[(y^{(p)})^2\right] - 2\mathbb{E}_Q\left[y^{(p)}\right]\mu_p + \mu_p^2}{2\sigma_p^2}\right) \\
&- \left(\frac{\mathbb{E}_Q\left[(z^{(r)})^2\right] - 2\mathbb{E}_Q\left[z^{(r)}\right]\nu_r + \nu_r^2}{2\tau_r^2}\right) \\
&= \underset{\theta'}{\operatorname{argmin}} \sum_{r=1}^{R}\sum_{p=1}^{P} \log \frac{1}{\sigma_p \tau_r} - \left(\frac{\mathbb{E}_Q\left[(y^{(p)})^2\right] - 2\mathbb{E}_Q\left[y^{(p)}\right]\mu_p + \mu_p^2}{2\sigma_p^2}\right) - \left(\frac{\mathbb{E}_Q\left[(z^{(r)})^2\right] - 2\mathbb{E}_Q\left[z^{(r)}\right]\nu_r + \nu_r^2}{2\tau_r^2}\right)
\end{aligned}
$$

Derivatives with respect to μ_p, ν_r, σ_p^2 and τ_r^2

$$
\begin{aligned}
\frac{\partial Q}{\mu_p} &= \sum_{r=1}^{R} \frac{2\mu_p - 2\mathbb{E}_Q\left[y^{(p)}\right]}{-2\sigma_p^2} = \sum_{r=1}^{R} \frac{2\mu_p - 2\mathbb{E}_Q\left[y^{(p)}\right]}{-2\sigma_p^2} = 0 \\
&\sum_{r=1}^{R} \frac{\mu_p - \mathbb{E}_Q\left[y^{(p)}\right]}{\sigma_p^2} = 0 \Rightarrow \mu_p = \frac{1}{R}\sum_{r=1}^{R} \mathbb{E}_Q\left[y^{(p)}\right]
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial Q}{\nu_r} &= \sum_{p=1}^{P} \frac{2\nu_r - 2\mathbb{E}_Q\left[z^{(r)}\right]}{-2\tau_r^2} = \sum_{r=1}^{R} \frac{2\nu_r - 2\mathbb{E}_Q\left[z^{(r)}\right]}{-2\tau_r^2} = 0 \\
&\sum_{r=1}^{R} \frac{\nu_r - \mathbb{E}_Q\left[z^{(r)}\right]}{\tau_r^2} = 0 \Rightarrow \nu_r = \frac{1}{P}\sum_{p=1}^{P} \mathbb{E}_Q\left[z^{(r)}\right]
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial Q}{\sigma_p} &= \sum_{r=1}^{R} -\frac{1}{\sigma_p} + \frac{\mathbb{E}_Q\left[(y^{(p)})^2\right] - 2\mathbb{E}_Q\left[y^{(p)}\right]\mu_p + \mu_p^2}{\sigma_p^3} = 0 \\
&\frac{1}{\sigma_p^3}\sum_{r=1}^{R} \mathbb{E}_Q\left[(y^{(p)})^2\right] - 2\mathbb{E}_Q\left[y^{(p)}\right]\mu_p + \mu_p^2 = \frac{R}{\sigma_p} \\
&\sigma_p^2 = \frac{1}{R}\sum_{r=1}^{R} \mathbb{E}_Q\left[(y^{(p)})^2\right] - 2\mathbb{E}_Q\left[y^{(p)}\right]\mu_p + \mu_p^2
\end{aligned}
$$

$$\frac{\partial Q}{\tau_r} = \sum_{p=1}^{P} -\frac{1}{\tau_r} + \frac{\mathbb{E}_Q\left[(z^{(r)})^2\right] - 2\mathbb{E}_Q\left[z^{(r)}\right]\nu_r + \nu_r^2}{\tau_r^3} = 0$$

$$\frac{1}{\tau_r^3}\sum_{p=1}^{P}\mathbb{E}_Q\left[(z^{(r)})^2\right] - 2\mathbb{E}_Q\left[z^{(r)}\right]\nu_r + \nu_r^2 = \frac{P}{\tau_r}$$

$$\tau_r^2 = \frac{1}{P}\sum_{p=1}^{P}\mathbb{E}_Q\left[(z^{(r)})^2\right] - 2\mathbb{E}_Q\left[z^{(r)}\right]\nu_r + \nu_r^2$$