1 Error Model for Mass Spectrometry

1.1 Context

In order to understand the Nf-κB signaling pathway of the cell a good model of the process has to be obatined. The state of the art aproche is used which is to define a structure of the model with certain parameters left to be fitted such that the model agrees with the messured data. This structure is generated by pure thought and prior knowleg of the systems. A vital part of this pursuite is to determine the functional dependencies of values in the model and there degree of freedome represented by the number of parameters that can adjust these dependencies. It is of high interesst to limit this number of paramters to a minimum that just allows the model to fit all realistic systems of the kind it is suppose to model but not more, as it could get "overfitted" which results in unrealistic predictions.

Since messured data is not free of noise this noise has to be accounted for when trying to fit the model. Thus the model should not predict the messured data exactly but the messured data should be destributed arround the prediction just as the messurment is distributed arround the real values due to the noise of messurment. Since these destributions are not know they have to be fittet as part of the model. The right type of destribution has to be picked and the parameters fittet such that it represents the destribution of the messurements arround the prediction and therefor hopefully the real value.

With the error model obtained we can calculate the probability of a certain messurent to be taken and hence the porobability to messure a set of certain values like the messurments already taken. This liklyhood to messure the existing data will be used as the goodness of fit. Thus fitting the selected model means trying to find a set of parameters for the model that maximizes this likelyhood.

1.2 Goal

Here we want to analyze the mechanics of the procedure to messure the data in order to obtain an understanding of the noise which is produced relative to the real values. We hope to derive a general functional depedency of the messured values to their errors with only a few paramters to be fitted. Experirience in this field of science has shown that such mechanistic approches ussually do not work perfectly due to the lack of accurate knowlege about the underlying process. Henc alternative modles will be worked out and compared by means of ACI, AIC$_C$, BIC, Analysis of Variance, Signal-to-Noise ratio, Cross-Validation and Minimum Description Length.

If sufficent data is provided we will also compute Shapiro-Wilk tests and the like to analyse the distribution of error arround an established prediction, e.g. multiple messurments of the same dilution of a certain protein. If there is more than one of such data sets we could also analyze the errors dependency to properties of the protein as size or lipophilicity.

1.3 Mechanic Description

1.3.1 Cell to Lysat

The cells which ought to be analyzed has a certain volume V_C and conatins a descrete number N_C of the protein of interest. This number can also be represented by the concentraion of the protein in the cell C_C by

$$N_C = V_C \cdot C_C$$

This cell lives in a culture of n mostly equal cells which is lysated to aquire the lysate with volume V_L, number of proteins $N_L = N_C \cdot n$ and an concentration $C_L = \frac{N_L}{V_L}$ which is the average of the protein concentrations of the individual cells of the culture. In the analyzis of a constant state this averaging can be an advantage as it corrects for errors due to biological diversity and absorbs some deviations of outlyers. But in the observation of a dynamic time dependent process it could be source of a major falsification of the data. The stimulus given to the culture does not reach each cell at an equal momemt of time and the cells probably do not react equally resulting in a divers set of reaction curves with peaks at different times and maybe even completly different shapes. An averaging at each individual point of time over all the cells would result in a reaction curve that is not representative for any typical cell reaction. Another uncertenty is weather intercellular communictaion can result in long forced delays of the muessured process in some cells. However the resulting deviation of the messured curve to the true cell respons is hard to compute and has to be ignored for now. As the process which is to be analyzed does only occure over several hours and is messured in timestepps of 30 minutes to 2 hours we hope be mimally effected by the diversity of respons in time.

1.3.2 Lysat to Detection

However all the folling steps to generate the resulting messurment can be modeled in a lot easyer fassion and will be adressed now in a much more satisfying manner. As the protein travels from the lysate through the filters, ionisation, splicing and finally detection by a fraction of focus it has probabilty of $1 - p$ to disappear and a probability p to be detected which is equal for each copy of the same protein. Hence we observe a Bernoulli trial and the number or cencentration of detected proteins would be binomialy distributed. As we process a very high number of proteins this destribution converges to a normal destribution and additional noise of messurement and lack of resultion makes them indestinglishable. Thus this error can be modeled with a addidative normal destribution. The random variable of number of detected proteins N_D can be discriped as

$$N_D = N_L + \varepsilon_L$$

where $\varepsilon_L \sim \mathcal{N}(0, \sigma)$ is the normally distributed error due to random loss of proteins.