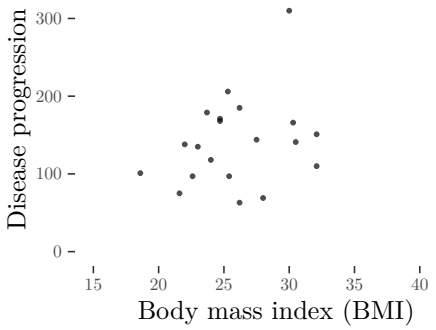
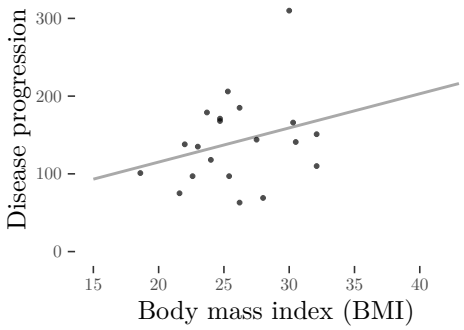
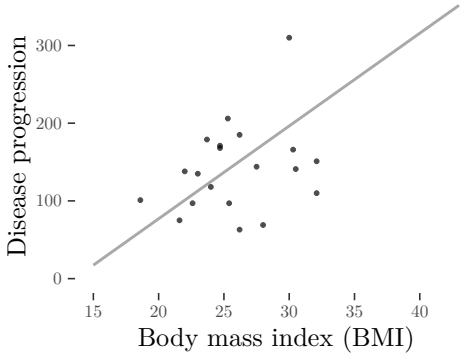
Lectio praecursoria – English subtitles

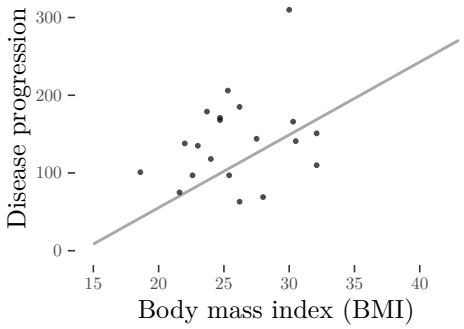
Juho Piironen

24 May 2019

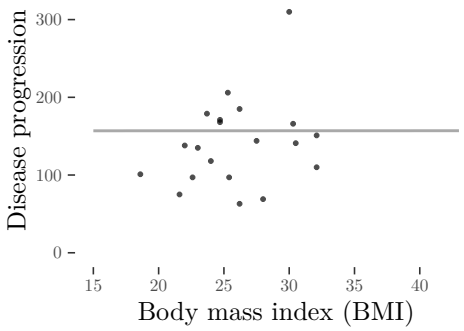
Figure / Slide	Subtitles
Bayesian Predictive Inference and Feature Selection for High-Dimensional Data	The aim of this talk is to give a brief overview and idea of the research topic of the thesis.
Bayesian Predictive Inference and Feature Selection for High-Dimensional Data	Let us begin with what the word "Bayesian" stands for.
	The plot on the left shows an example dataset collected from type 2 diabetes patients. Visualized are a quantitative measure of disease progression after one year baseline (y-axis) versus the body mass index (x-axis). (The units of the disease progression are unimportant; it is sufficient to know that high values indicate more rapid progression).
	To study the statistical relationship between the two variables, we can fit a simple statistical model $$y = \beta x + \alpha$$ to these data. The plot on the left shows the line that best fits the data (loosely speaking, in terms of being as close to all the points as possible). Based on this line, it appears that the two variables truly are statistically related, meaning that patients with higher body mass index tend to have more rapid disease progression.



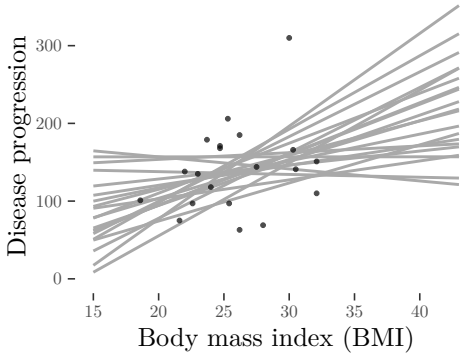
However, since we have very limited data, we have uncertainty about where the optimal line goes. It could also look like this....



...or this...

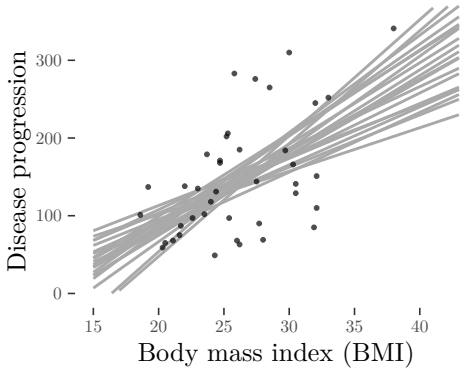


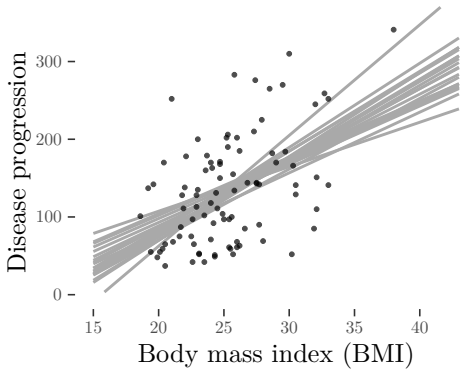
...or this.

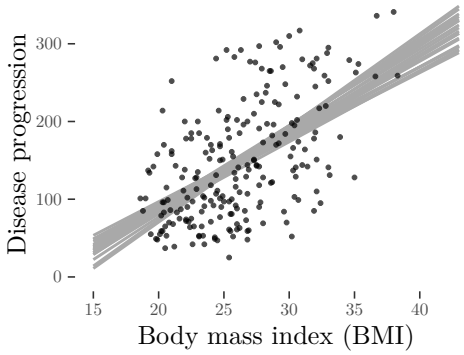


In fact, there are lots of lines which would be plausible (likely) given these data.

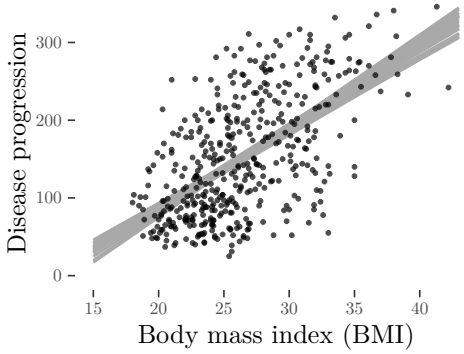
This is in a nutshell what Bayesian inference is about: quantifying uncertainty.



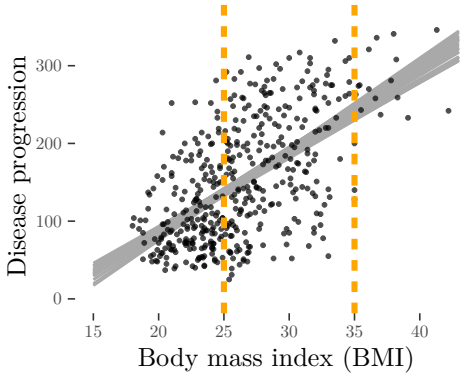




We can reduce the uncertainties by making more observations. Once we collect more data, we gradually get a better understanding where the true line goes.

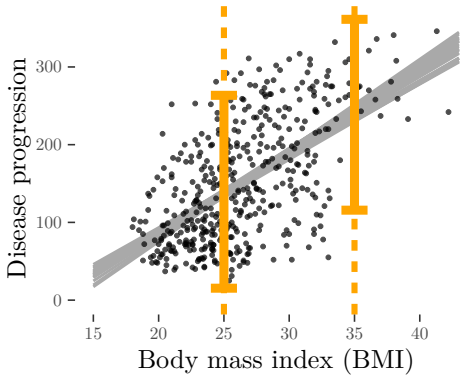


After a certain number of observations, we already have a pretty good idea of the true statistical relationship between the two variables.



In addition to providing us with understanding about the relationships between different variables, the statistical models have the capability of making predictions.

For example, if we compare a person with severe overweight (BMI of 35) to a normal weight person (BMI of 25), we can easily calculate what is difference between the expected disease progression.



This being said, it is important to understand that the expectations are not the whole story.

For example, we look at the intervals containing 95 percent of the patients, we see that it is well possible that a person with normal weight can easily have more rapid disease progression. This variation is explained by other factors than the BMI.

No figure.

Great. Now we should have some idea of what a simple dataset and statistical model might look like, and hopefully we also have some sense of what Bayesian inference refers to.

No figure.

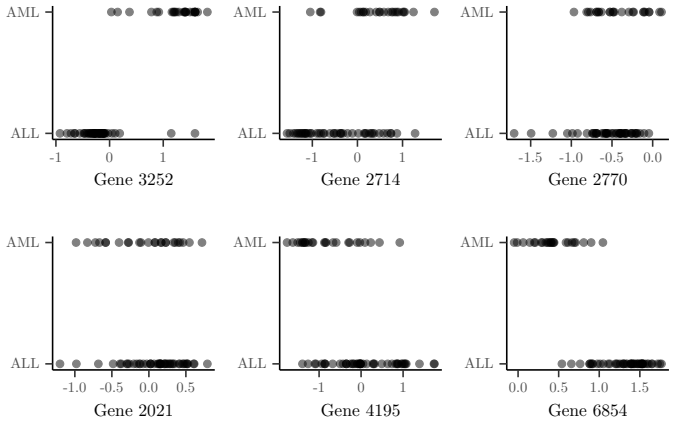
Let us then look at what we mean by "high-dimensional" data.

We often want to model some target y with many features $x_1, \ldots, x_d$. When the number of features d is large, the data are said to be *high-dimensional*.

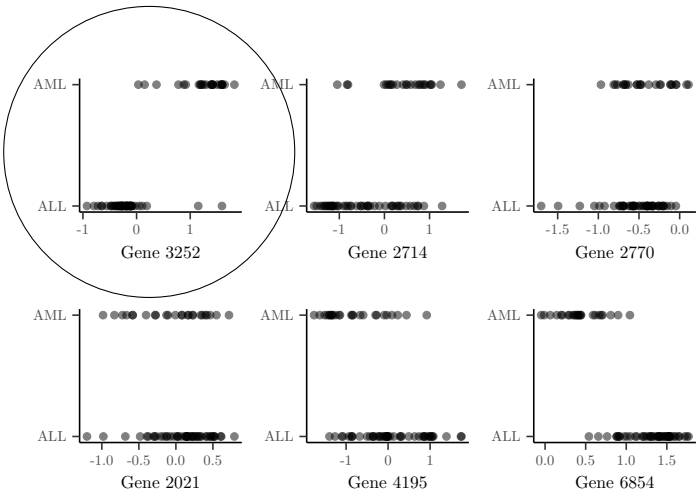
Example: data collected from $n = 72$ leukemia patients, where

y : ALL (acute lymphoblastic leukemia), or AML (acute myeloid leukemia)

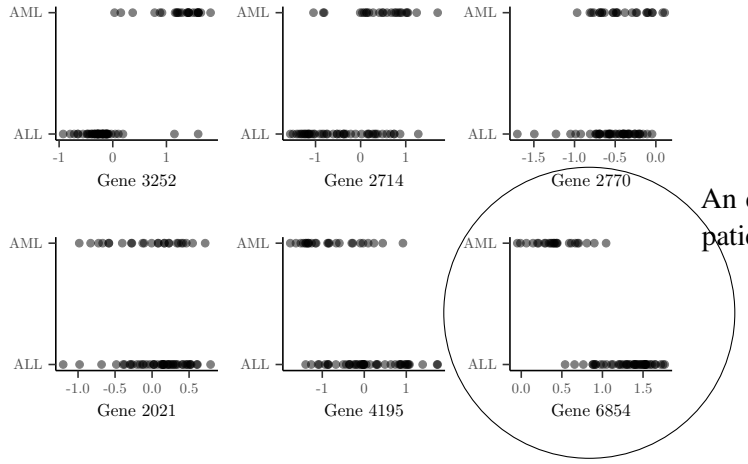
$x_1, \ldots, x_d$: Gene expressions for $d = 7129$ genes



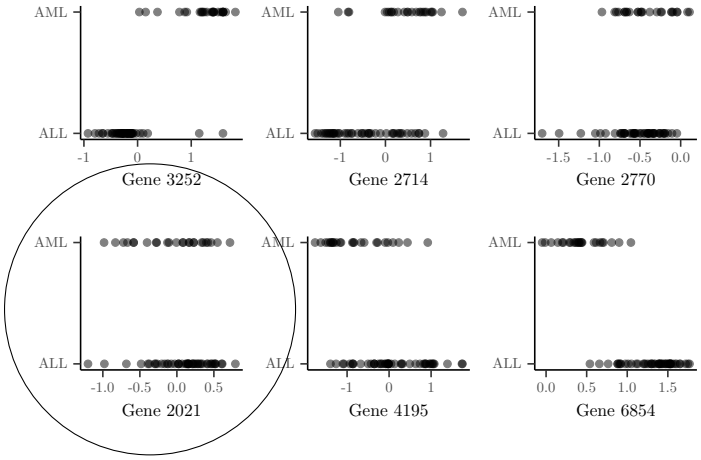
Here are visualized some example gene expressions for the leukemia dataset.



An example of a gene with higher expression for *AML* patients.



An example of a gene with higher expression for *ALL* patients.



An example of a gene with about similar expression levels for both groups.

No figure.

Alright, now we have an idea of what high-dimensional datasets are. Let's then move on to see what type of problems the thesis studies.

Bayesian Predictive Inference and Feature Selection for High-Dimensional Data

The thesis focuses on predictive inference and feature selection. What is meant by these?

No figure.

Prediction:
"Can we classify the cancer of a patient given only his or her gene expressions?"

- Predictions are often intrinsically valuable.
- Predictive accuracy has also value simply in figuring out whether the features (genes) can explain the variation in the class (cancer type).

Feature selection:
"Which genes are relevant for the prediction?"

- Feature selection improves the interpretability and aids understanding the underlying phenomenon.
- Feature selection can also reduce future costs if there is a price associated with predicting with many features.

There are several challenges related to these tasks.

Statistical challenges, e.g.:

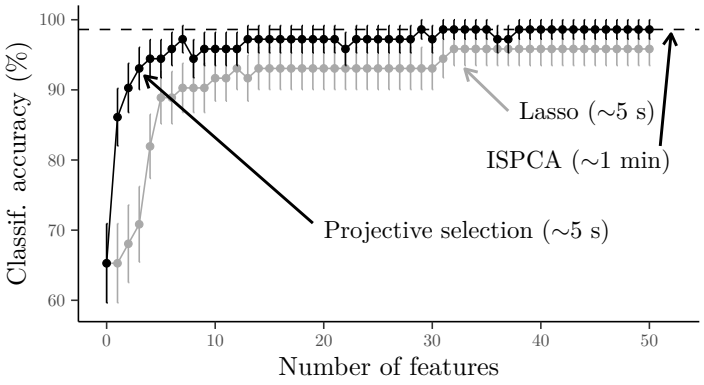
- Model parameters are non-identifiable without further assumptions.
- There are plenty of uncertainties due to the high-dimensional feature space and small sample sizes.

No figure.

Computational challenges, e.g.:

- Making inferences on the model parameters using all the features is computationally heavy.
- In feature selection it is impossible to go through all the possible feature combinations. One must resort to heuristics instead.

The aim of the thesis: develop accurate but computationally feasible prediction and feature selection methods for these datasets.



Here is an illustration of the prediction and feature selection for the Leukemia example.

With method proposed in the thesis (ISPCA) we can get predictions that are more accurate than those obtained with Lasso (a popular non-Bayesian alternative). The computation time (about a minute) is acceptable increase to that of Lasso (about five seconds). The downside is that the model uses all features.

We can simplify the ISPCA solution using the projective selection, which is computationally as efficient as Lasso, and yields a superior tradeoff between accuracy and the number of features used.

In a nutshell: with the techniques proposed in the thesis, one can find more accurate and simpler models – in a computationally efficient manner. Although the examples during this talk have been with medical datasets, the proposed methods are generic and can be applied to datasets from various fields.

Bayesian Predictive Inference and Feature Selection for High-Dimensional Data

We hopefully now have some idea of what the thesis is about. It is worth pointing out, that even though here we have used medical datasets as examples, the methods are generic and applicable to datasets from various fields.