

# Differential Privacy and Synthetic Data Generation using PrivBayes

Jan Reiter Sørensen



AALBORG  
UNIVERSITY



# USER GUIDE

- REMEMBER TO DELETE THIS SLIDE BEFORE YOU FINISH YOUR PRESENTATION

## AAU POWERPOINT TEMPLATES

When you open PowerPoint, you can choose between two templates in widescreen (16:9) with the AAU logo in either Danish or English.

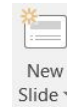
You can also download the templates at: [www.design.aau.dk/skabeloner/powerpoint](http://www.design.aau.dk/skabeloner/powerpoint)  
You can also download the AAU PowerPoint presentation for inspiration. We will make sure that the AAU PowerPoint presentation available for download is updated with the latest figures and information. When you download the AAU PowerPoint presentation, you can simply delete the slides you do not wish to use in your own presentation.

## FONTS

In PowerPoint presentations we use AAU's secondary font Arial.

## INSERT A NEW SLIDE

1. Click **Home**
2. If you want to duplicate a slide layout, select the slide you wish to duplicate and click the top half of the **New slide** button. Click the bottom half of the New slide button to view the selection of layouts available with the AAU design.



## GRIDLINES AND GUIDES

You can use gridlines and guides to help you align and place objects and improve the overall design and layout of your presentation. To view gridlines and guides:

1. Select **View**
2. Select **Gridlines** and/or **Guides**

**Quick tip:** Press **Alt + F9** to show gridlines

## TEMPLATE COLOURS

You can choose from a range of colours for your backgrounds and graphs.

Right-click the shape you wish to fill with colour and select the paint bucket (Shape fill)

## IMAGE SIZE

Remember to use high-quality images for your layouts. Avoid making small images larger, the images become very poor quality and appear grainy and unprofessional.

You can search for and download high-quality AAU images in Skyfish.

Naturally, many high-quality images will increase the overall size of your PowerPoint file. If your PowerPoint file becomes too big, you can resize it when you save your presentation, select **Save as**, select a destination folder by clicking Browse, click the **Tools** button and select **Compress pictures**. Always select Screen (150 ppi) or above.

## INSERT IMAGES

For layouts containing picture placeholders: Click on the icon in the placeholder or select **Insert** and click **Pictures**



The text in the placeholder will not appear in your final presentation

## CROP IMAGES

1. Select the image you want to crop, (*rightclick on image and click on the **Crop** icon*)



2. To scale the image (resize the image proportionally), hold the **SHIFT** key while you drag one of the four corner handles

3. Right-click the image and select **Send to back**

**Quick tip:** If you delete the image and replace it with a new image, this image may appear in front of the text and graphics. If this happens, right-click the image and select **Send to back**

## MORE INFORMATION

You'll find more information at [www.design.aau.dk/skabeloner/powerpoint](http://www.design.aau.dk/skabeloner/powerpoint)

# Agenda

- Problem statement
- Differential privacy
- Differentially private mechanisms introduced in this project
  - The smoothed histogram mechanism
  - The Laplace mechanism
  - The exponential mechanism
  - PrivBayes
- Investigating synthetic data by PrivBayes
  - Generating some data and sanity check for generating differentially private Bayesian networks
  - The mixture parameter
  - Overall privacy and power of Pearson's test
- Conclusion

# Problem Statement boiled down

1. How are differentially private methods for synthesizing data defined and explained using a consistent and precise mathematical language?
2. How are differentially private synthesizers such as the smoothed histogram mechanism and PrivBayes implemented?
  - a. In practice
  - b. For instance using R?
3. What is the relationship between the privacy and the utility of synthetically generated data?

*“Differential privacy is a mathematical property of synthetic data generation methods, and rigorously showing that a synthesizer satisfies differential privacy needs thorough mathematical reasoning. Not all differentially private methods are introduced in a rigorous setting, and neither are they built on the same mathematical terminology. This leads to the following question. How are differentially private methods defined and explained using a consequent and precise mathematical language? There are several methods for synthesization of differentially private data such as the smoothed histogram mechanism and PrivBayes, but they are quite theoretical. How can these mechanisms, if possible, be implemented in practice using a programming language such as R? Furthermore, there is an inherent trade-off between the privacy and the utility of synthetically generated data. What is the relationship between the privacy and the utility of synthetic data?”*

# Differential Privacy

- The relative change of the conditional distribution with respect to a unit change in the original dataset is bounded by  $\exp(\alpha)$ .
- This guarantees a certain level of privacy even in the presence of outliers.
- For  $\alpha$  close to zero the conditional distributions turn indistinguishable.

$$\sup_{\substack{B \in \mathcal{B}(\mathbb{R}^{k \times d}) \\ (x, x') \in N(X)}} \frac{Q(B|X = x)}{Q(B|X = x')} \leq \exp(\alpha)$$

$$\frac{Q(B|X = x)}{Q(B|X = x')} \in [\exp(-\alpha), \exp(\alpha)]$$

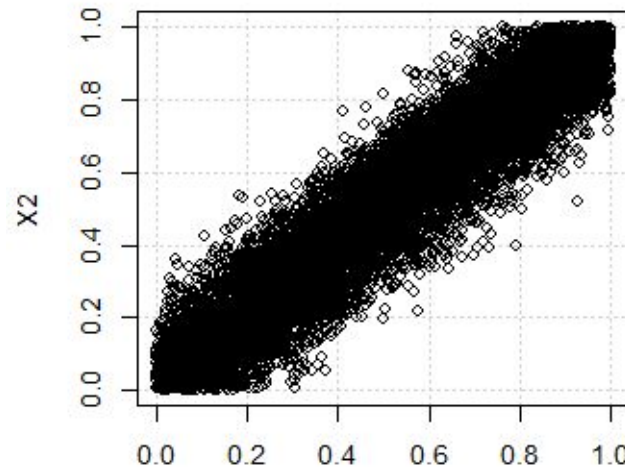
# The smoothed histogram mechanism

- Generates a histogram on the basis of a dataset.
- Smoothing is applied in order to introduce differential privacy.

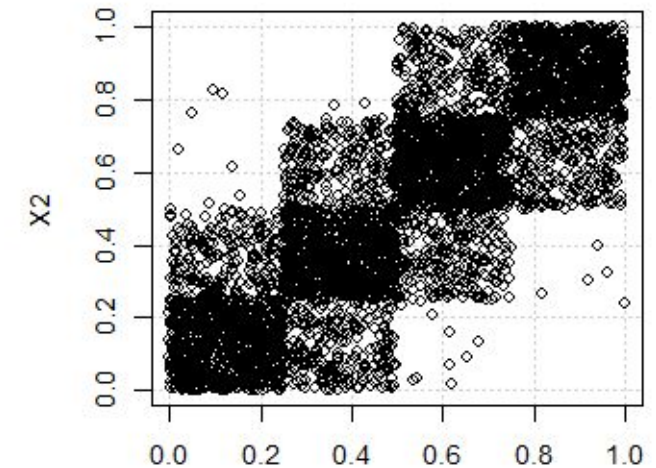
$$\alpha = k \ln \left( \frac{(1 - \delta)m}{n\delta} + 1 \right)$$

- Can easily be applied to any dataset consisting of numeric or categorical variables, **but scales poorly due to the curse of dimensionality.**

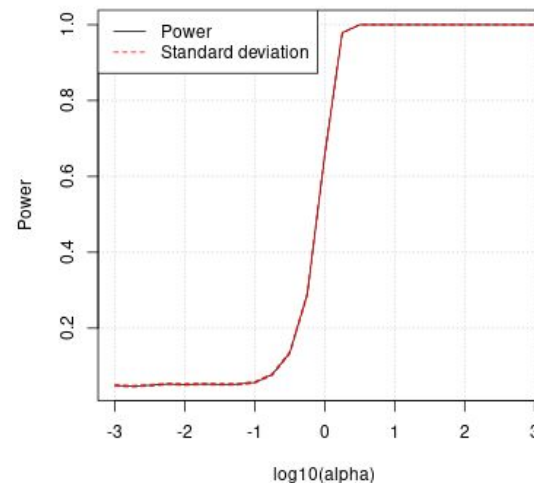
Original dataset X



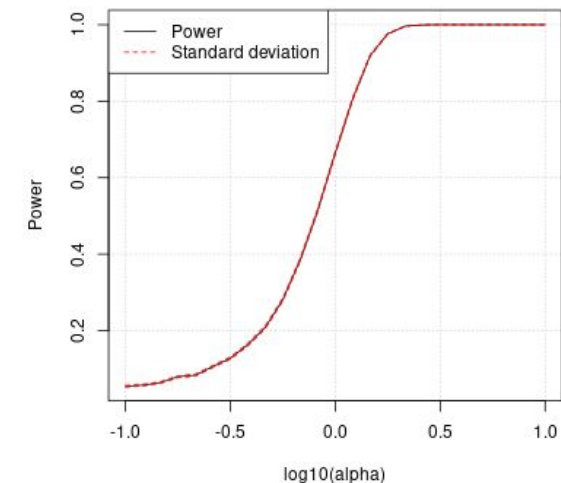
Sampled dataset Z



Power of Spearman rank test for correlation

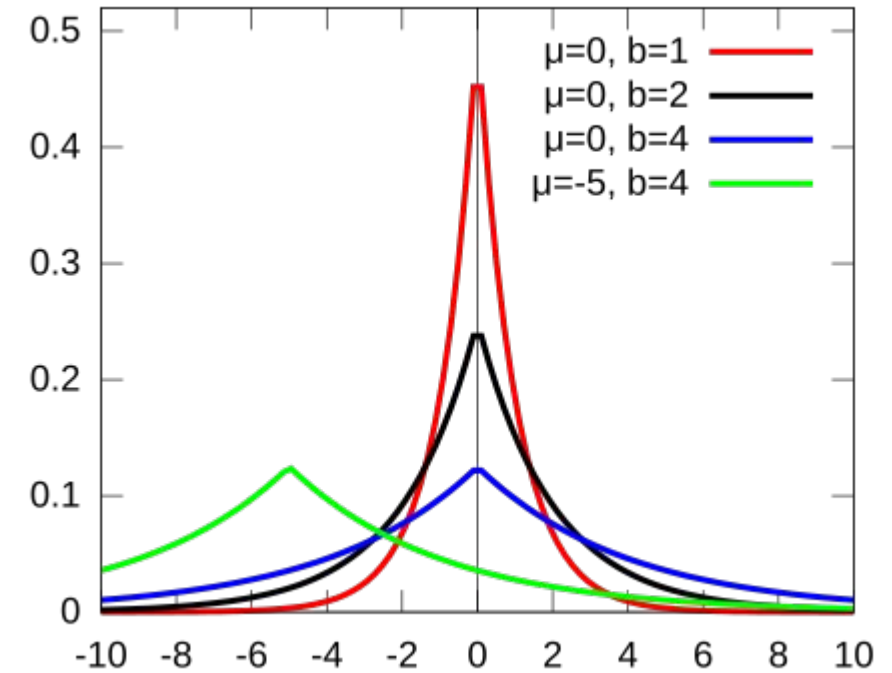


Power of Spearman rank test for correlation



# The Laplace mechanism and the exponential mechanism

- The Laplace mechanism:
  - Add Laplace distributed noise to each observation.
  - The scale parameter depends on the sensitivity and the wanted level of differential privacy.
  - Method can only be applied to continuous variables
- The Exponential mechanism:
  - The function  $q$  scores a potential output to an input, and outputs are then sampled such that high scores are exponentially more likely to be picked than lower scores.
  - Can be applied to any type of data, but can be difficult to implement.
  - Postulated in the literature to be a foundational differentially private mechanism, meaning that **all differentially private mechanisms are variants of the exponential mechanism.**



$$g_x(r) \propto \exp(\alpha q(x, r)),$$

$$Q(B|X = x) := \int_B g_x(r) d\mu(r)$$

# PrivBayes

- Utilizes Bayesian networks.
- We can implement this using the exponential mechanism for generating Bayesian networks differentially private.
- The Laplace mechanism can be used for injecting noise into the conditional probabilities.

## Algorithm 4.0.1. (PrivBayes)

Let  $X : \Omega \rightarrow \mathbb{R}^{n \times d}$  with  $n, d \in \mathbb{N}$  be a database, let  $V = (X_1, X_2, \dots, X_d)$  be a random vector that is equal in distribution to each row of  $X$ . The PrivBayes mechanism is given by the following algorithm

1. Fit a  $k$ -degree Bayesian network

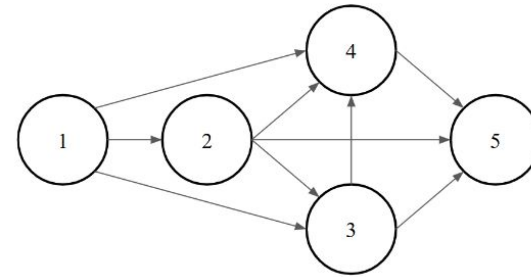
$$\mathcal{N} = \{(1, \Pi_1), (2, \Pi_2), \dots, (d, \Pi_d)\},$$

that resembles the conditional independence relations in  $V$  for some chosen  $k \in \mathbb{N}$ , using an  $\alpha_1$ -differentially private method.

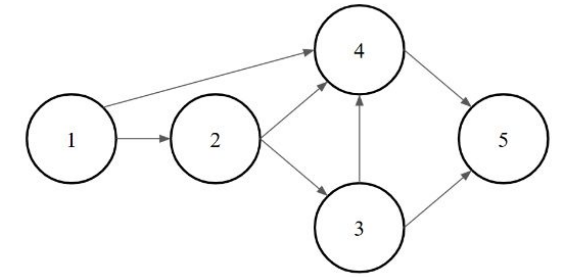
2. Estimate conditional probabilities  $\mathbb{P}(X_i | X_j, j \in \Pi_i)$  for all  $i = 1, 2, \dots, d$ , and inject noise using an  $\alpha_2$ -differentially private method.
3. Using the noisy conditional distributions, assemble a noisy joint distribution, and sample a synthetic dataset.



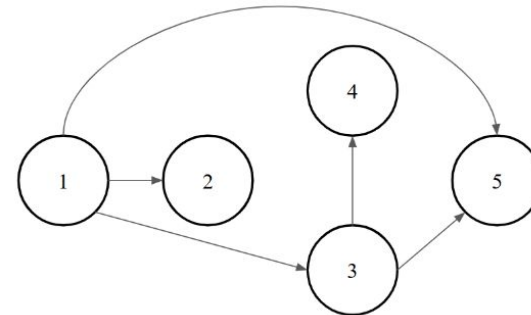
# Generating some data and sanity checking the Bayesian network generating mechanism

$$\begin{bmatrix} 1 & 0.6 & 0.5 & 0.4 & 0.3 \\ 0.6 & 1 & 0.6 & 0.5 & 0.4 \\ 0.5 & 0.6 & 1 & 0.6 & 0.5 \\ 0.4 & 0.5 & 0.6 & 1 & 0.6 \\ 0.3 & 0.4 & 0.5 & 0.6 & 1 \end{bmatrix}$$


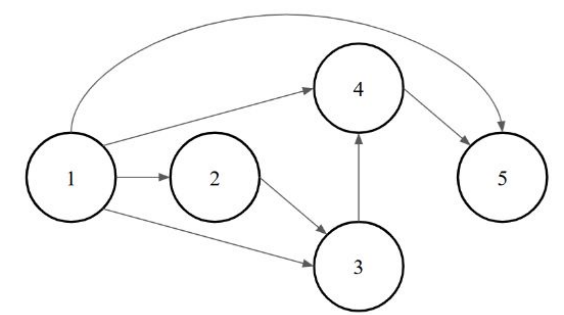
(a) Generated under 50-differential privacy.



(b) Generated under 5-differential privacy.



(c) Generated under 0.5-differential privacy.



(d) Generated under 0.05-differential privacy.

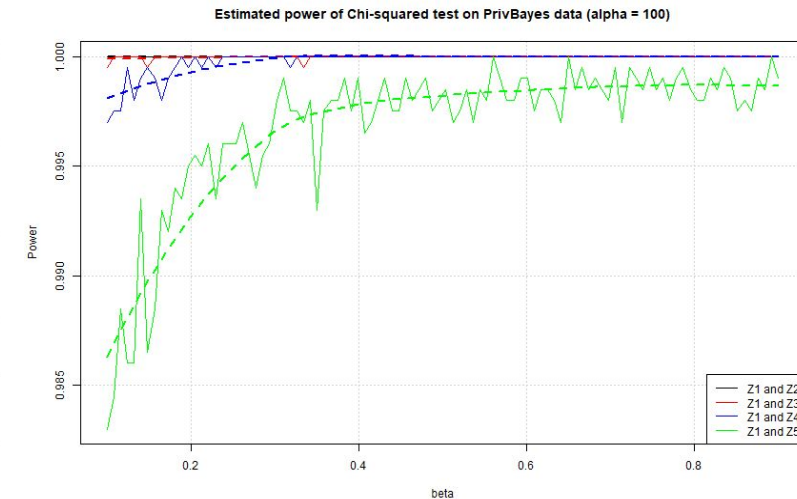
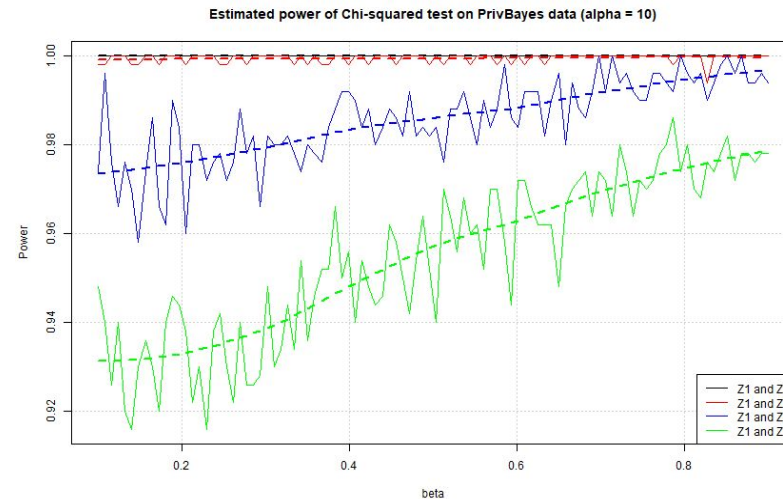
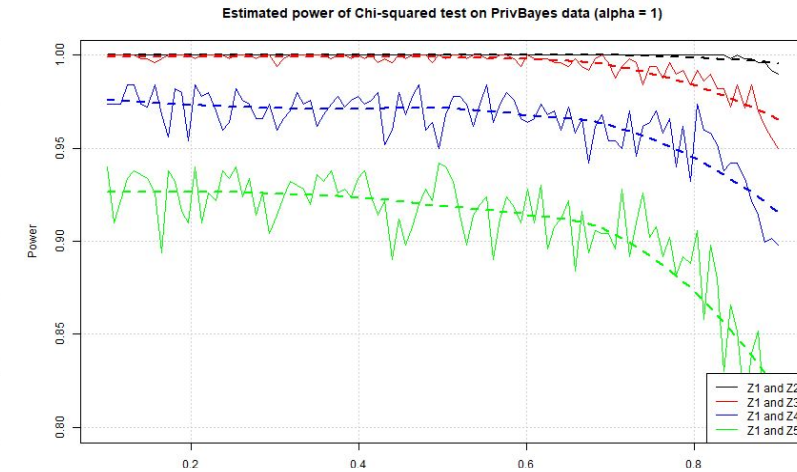
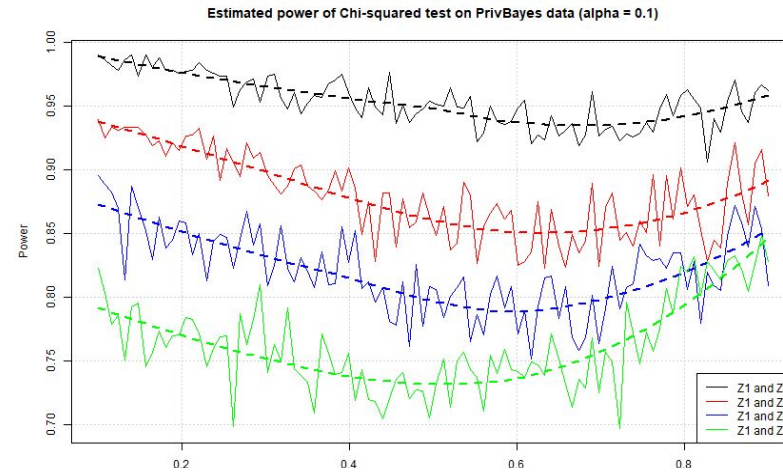
# The mixture parameter

$$\alpha := (\alpha_1 + \alpha_2)$$

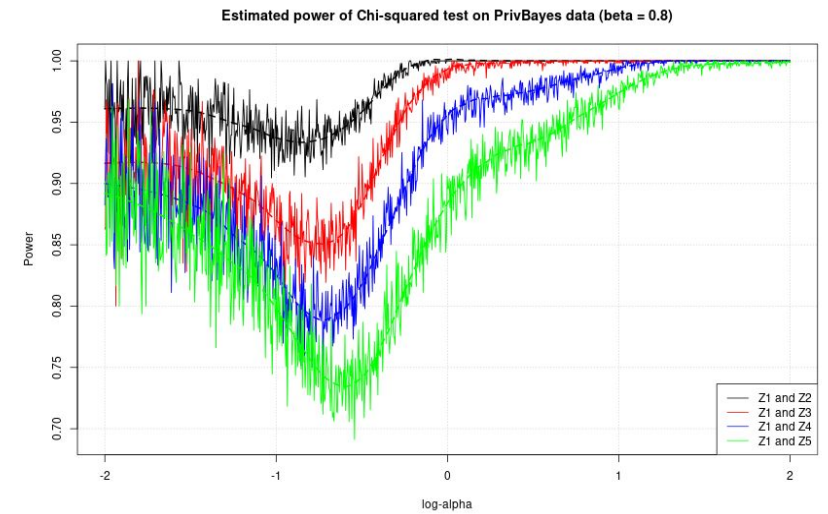
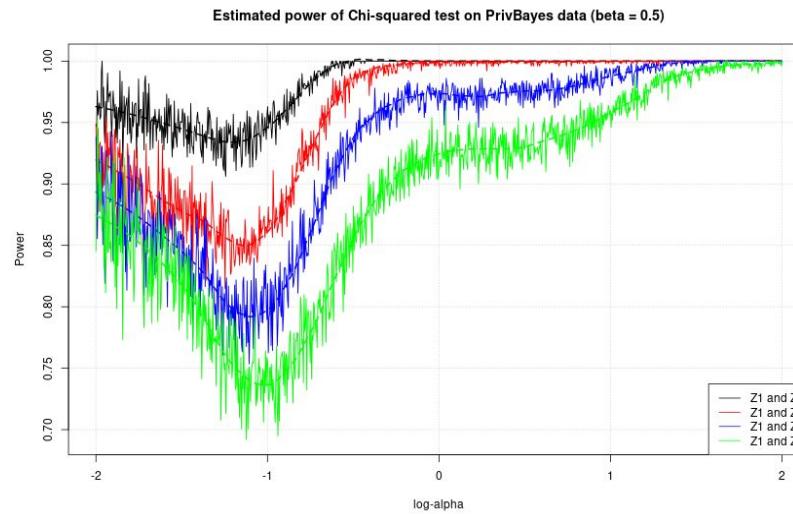
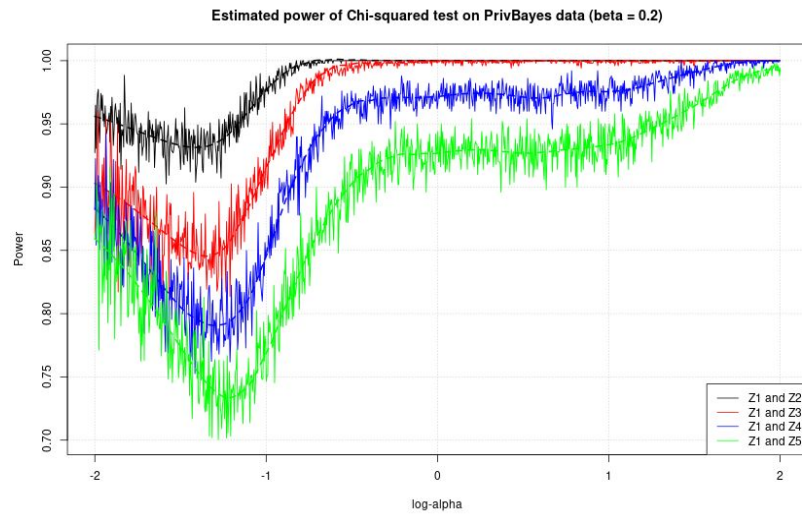
$$\beta := \frac{\alpha_1}{\alpha} \quad \beta \in [0, 1]$$

$$\alpha_1 = \beta\alpha$$

$$\alpha_2 = (1 - \beta)\alpha$$



# Overall privacy and power of Pearson's test



# In conclusion

- We saw that the relationship between the mixture parameter and power depended on the overall privacy budget.
  - Extreme privacy: Either high or low beta
  - Slightly less privacy: Important to not have too much privacy on the conditionals.
  - Low privacy: Important to not have too much privacy on the Bayesian network generating mechanism.
- We saw that the power decreased as the overall privacy budget decreased, that is a higher level of privacy implied lower power. Comparing PrivBayes to the histogram method, we saw a much better privacy-utility trade off for PrivBayes data.



# Acknowledgements

- For supervision
  - Martin Bøgsted and Heidi Søgaard
- For good advice and considerations
  - Jakob Skelmosø and Rasmus Rask



# Questions