Unsupervised automatic dataset repair

James Allingham supervisors: Prof Zoubin Ghahramani & Dr Christian Steinruecken
University of Cambridge

Missing Data

What is missing data?

π	0	1	?	1	NaN	0.1
e	0	0	?	?	∞	5
π	0	?	?	3	4	5
?	?	?	?	4	?	0.1
?	?	?	?	?	?	?
π	0	0	?	0.23	?	5
e	0	?	?	1.1	?	?

Types of missing data: Given a dataset $\mathbf{X}$ and a mask $\mathbf{M}$ describing which elements of $\mathbf{X}$ are missing, we can classify missing data as:

- missing completely at random (MCAR): $p(\mathbf{m}_i) \perp\!\!\!\perp p(\mathbf{x})$,
- missing at random (MAR): $p(\mathbf{m}_i) \perp\!\!\!\perp p(\mathbf{x}_i)$, or
- not missing at random (NMAR): $p(\mathbf{m}_i) \not\perp\!\!\!\perp p(\mathbf{x})$

where $\mathbf{x}$ is a row in $\mathbf{X}$, $\mathbf{x}_i$ is the value at column i of that row, and $\mathbf{m}_i$ is the mask for that value (Little and Rubin, 2002).

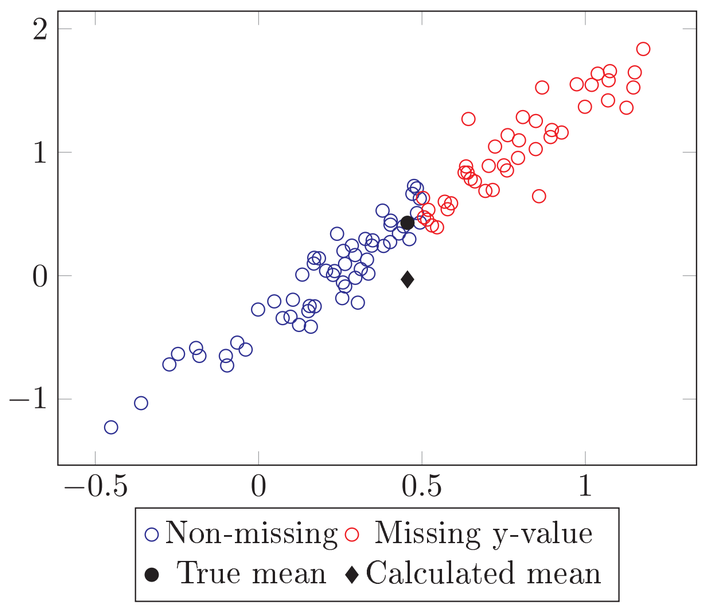
Causes for missing data: There are many reasons that data might be missing or corrupted:

- faulty sensors e.g. voltage is not measured for currents larger than 1A,
- unanswered survey questions e.g. people declining to give their salary,
- data corruption e.g. HDD failure corrupting a file,
- *etcetera.*

But why do we care about missing data?

The problem

How can we analyse data which has missing values? For example, can we calculate the mean of the dataset shown below? Simple strategies such as ignoring examples with missing values or replacing missing values with the observed mean can bias our results.



A solution

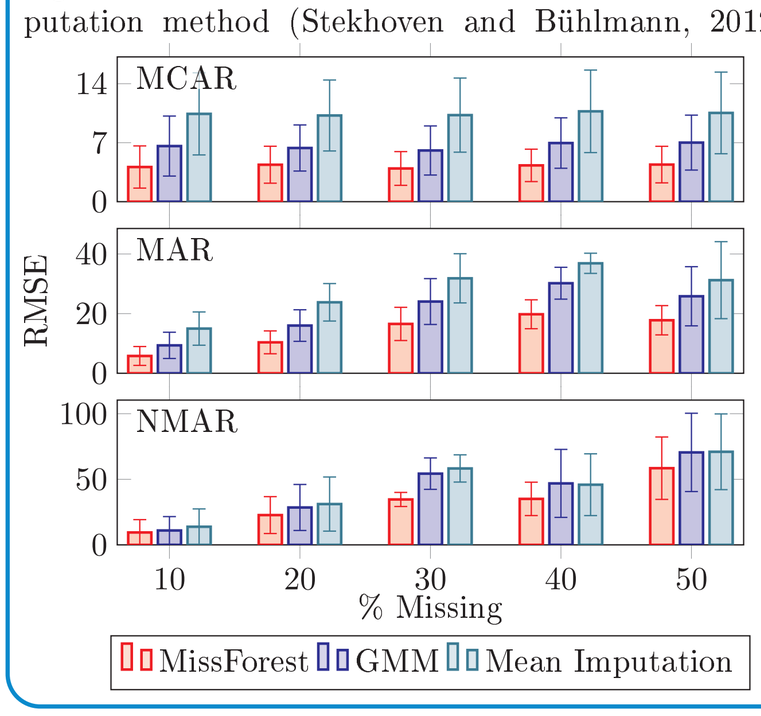
One solution is to repair the dataset by **imputing** the missing values. My research centered around using *probabilistic graphical models*, such as Gaussian Mixture Models (GMMs), and *stochastic processes*, such as Dirchlet processes, to learn the distribution for the data, despite the presence of missing values. After learning the distribution of the data it is possible to impute the missing values by taking the mode of the distribution or by sampling from the distribution. Using probabilistic machine learning methods has a number of advantages:

- they give us estimates for the uncertainty in our predictions,
- they provide a principled method for model comparison, and
- they allow us to perform *multiple imputation.*

However, probabilistic models are not state-of-the-art for imputation.

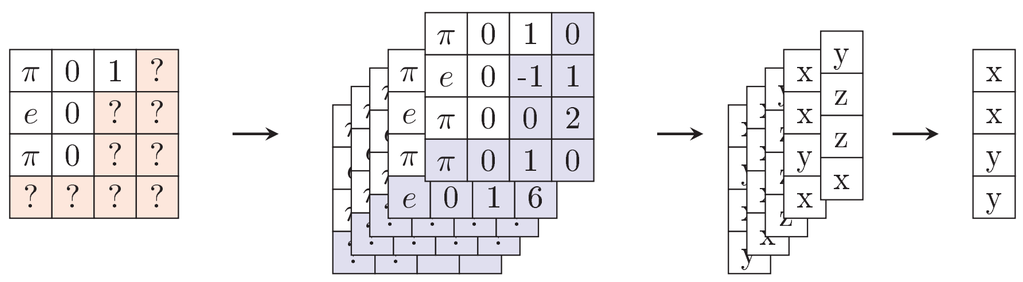
Results

Below is a comparison of imputation algorithms for Boston Housing with artificially introduced missing values. MissForest is a state-of-the-art imputation method (Stekhoven and Bühlmann, 2012).

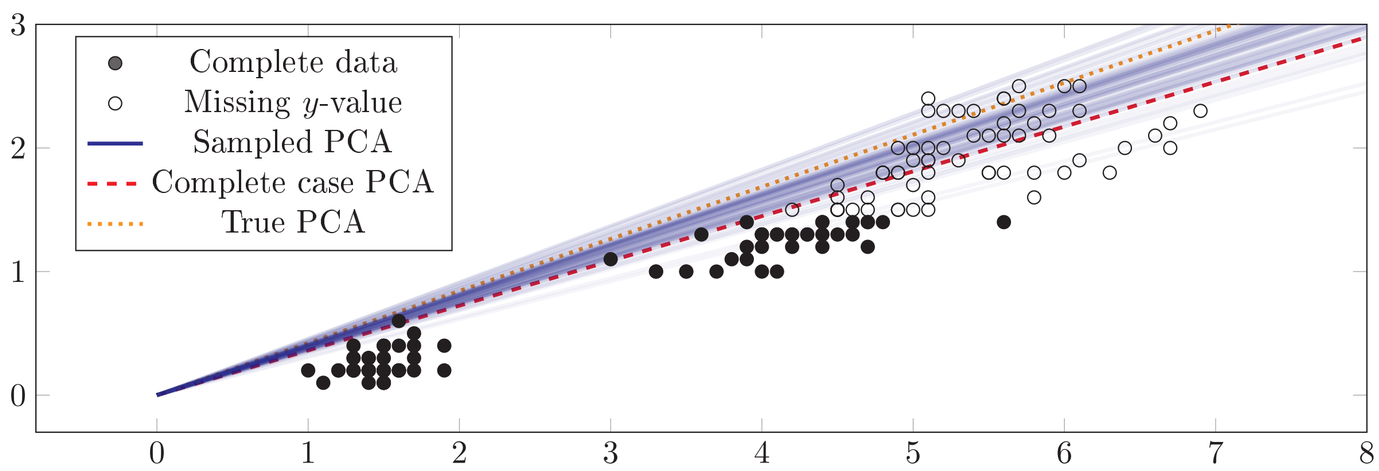


Multiple imputation

To highlight the uses of probabilistic models for imputation, lets take a closer look at multiple imputation! Here we impute a dataset N times by sampling from the distribution of the data. We then perform some analyses N times, and finally aggregate the analyses which gives us a robust final analysis with uncertainty estimates.



For example, what if we wanted to perform principle component analyses (PCA) on a dataset with missing values?



Future Work

Deep learning models have recently been shown to perform state-of-the-art imputation of missing data (Yoon et al., 2018). More specifically, Yoon et al. (2018) use generative adversarial networks to impute missing values. Other deep learning methods such as de-noising auto-encoders have been shown to perform well on imputation tasks.

However, deep learning methods typically do not give uncertainty estimates, allow for sampling from the distribution of the data, or provide principled model comparison.

That said, *probabilistic deep learning* is a growing field with lots of exciting research. Methods such as *auto-encoding variational Bayes* (aka variational auto-encoders) do have some of these properties, and have already been shown to work for imputation (Rezende et al., 2014).

Other research is aimed at getting uncertainty estimates from standard deep learning models such as CNNs (Gal and Ghahramani, 2016). These methods might allow us to have the best of both worlds – deep learning's raw modelling power, and probabilistic machine learning's uncertainty estimates!

Conclusion

Missing data is common in real-world datasets. On the other hand, imputation is an open problem with many areas for improvement. While probabilistic machine learning methods have some desirable qualities for repairing datasets with missing values, their performance is weaker than alternative approaches such as random forests, and deep learning models.

Probabilistic deep learning is an exciting direction for future research into missing data imputation! With deep learning models that produce uncertainty estimates, we could achieve state-of-the-art results while still having access to useful tools such as multiple imputation, and Bayesian model comparison.

Extra Information

Email: james.allingham@gmail.com **Code:** https://github.com/JamesAllingham/AutoImpute
Dissertation: https://jamesallingham.co.za/pdf/james_allingham_dissertation.pdf

References

Gal, Yarin and Ghahramani, Zoubin (2016). Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. URL https://arxiv.org/pdf/1506.02142.pdf.

Little, Roderick J. A. and Rubin, Donald B. (2002). *Statistical Analysis with Missing Data*. John Wiley & Sons, Inc., 2nd edition. ISBN 978-0-471-18386-0.

Rezende, Danilo J, Mohamed, Shakir, Wierstra, Daan and DeepMind, Google (2014). Stochastic Backpropagation and Approximate Inference in Deep Generative Models. URL https://arxiv.org/pdf/1401.4082.pdf.

Stekhoven, Daniel J. and Bühlmann, Peter (2012). MissForest – non-parametric missing value imputation for mixed-type data. *Bioinformatics*, **28**(1) 112–118. URL http://dx.doi.org/10.1093/bioinformatics/btr597.

Yoon, Jinsung, Jordon, James and Van Der Schaar, Mihaela (2018). GAIN: Missing Data Imputation using Generative Adversarial Nets. Technical report. URL http://medianetlab.ee.ucla.edu/papers/ICML{_}GAIN.pdf.