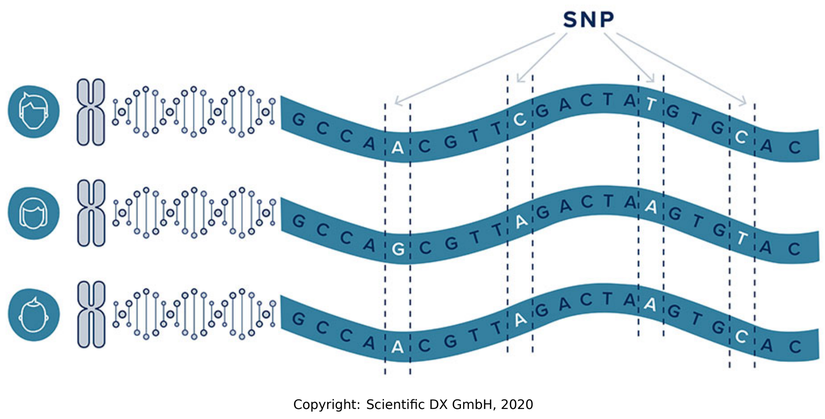
Genotype Imputation with Multi-label Random Forests

Ekaterina Antonenko and Jesse Read

Laboratoire d'informatique, École Polytechnique, IP Paris

Missing values in Single Nucleotide Polymorphism (SNP) data



- Single Nucleotide Polymorphism is a genomic variant at a single base position in the DNA.
- SNP data is used to study how the genome influences diseases and traits.
- Categorical encoding: 0 (dominant-dominant), 1 (dominant-mutant), 2 (mutant-mutant).
- Typically: data is high-dimensional $p \sim 10^5 - 10^6$ and low-sampled $N \sim 10^2 - 10^3$.
- Ordering is important due to linkage disequilibrium.
- Missing values occur due to external mechanisms $\Longrightarrow$ we study missing completely at random case.

Why data is missing?

- Errors in sensors.
- Errors in data processing.
- Combining different studies into one.

Why to impute missing data?

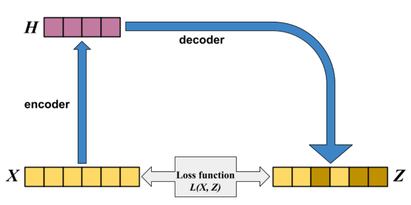
- Most off-the-shelf statistical and machine learning methods cannot handle missing values.
- Considering only instances with complete information can lead to a loss of necessary information and can yield a very poor or even empty dataset.
- Missing data itself might be of interest.

Imputation methods:

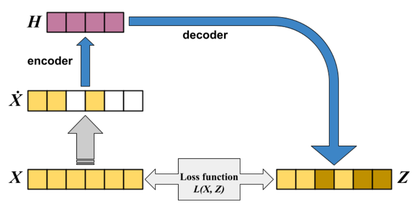
- Reference-based: use reference panels of high quality and sequencing covering (state-of-the-art for human data).
- Reference-free: use only data itself (when reference panels are not available, often this is the case for non-human data).

Reference-free methods		Limitations
Mode / mean / median	$\Longrightarrow$	Poor performance, don't use other features.
k Nearest Neighbors (kNN)	$\Longrightarrow$	It works, but can we do better?
Singular Value Decomposition (SVD)	$\Longrightarrow$	It works, but can we do better?
MissForest	$\Longrightarrow$	Too slow for high-dimensional data.
Denoising Autoencoders (SCDA)	$\Longrightarrow$	Require complete data for training.

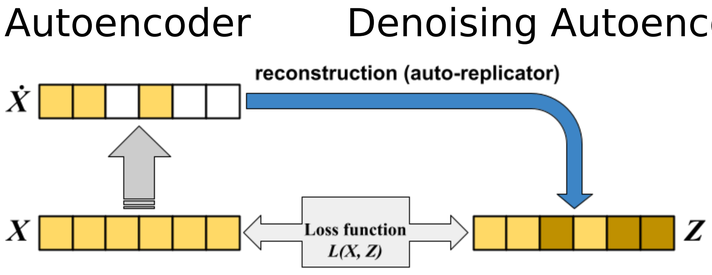
Our approach: Chains of Autoreplicative Random Forests (ChARF)



Autoencoder



Denoising Autoencoder

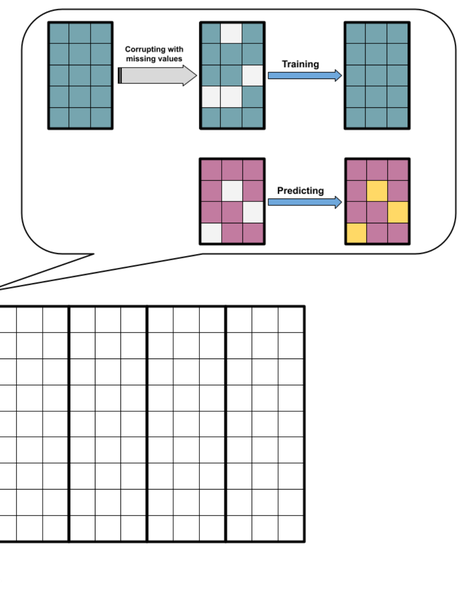


Autoreplicator

Why to use multi-label methods?

- Fewer parameters than neural networks (good for low-sampled data).
- No need for hidden layers.

BUT... Still need complete data for training. $\Longrightarrow$



Chains of autoreplicators

- One window of size Δ = training part with complete data + testing part with missing values.
- Chain of windows: on each step, stacking ν windows with already imputed values as additional features.
- Ensemble of chains: one forward chain, one backward chain, several random chains.

Results

Datasets used in experiments, p features, N samples:

Name	p	N
Maize	44,729	247
Eucalyptus	33,398	970
Colorado Beetle	34,186	188
Arabica Coffee	4,666	596
Wheat (Zuchtwert study)	9,763	388
Coffea Canephora	45,748	119

Gridsearch for hyperparameters Δ and ν:

Lighter color / higher accuracy

Δ: as expected, bigger fraction of missing values $\rightarrow$ smaller size of window $\Longrightarrow$ can be estimated theoretically, no need for search.
ν: may be different.

Comparison with other methods:

	0.01	0.05	0.1	0.2	0.3	0.01	0.05	0.1	0.2	0.3	0.01	0.05	0.1	0.2	0.3
	Maize					Eucalyptus					Coffea Canephora				
	(15, 1)	(15, 1)	(10, 1)	(5, 1)	(5, 1)	(10, 5)	(5, 10)	(5, 10)	(3, 10)	(3, 10)	(10, 1)	(10, 1)	(5, 1)	(5, 1)	(3, 1)
ChARF	**0.952**	**0.935**	**0.916**	**0.882**	**0.845**	**0.970**	**0.950**	**0.926**	**0.866**	0.810	**0.799**	**0.781**	**0.761**	0.731	0.717
kNN (5/10/10)	0.803	0.802	0.801	0.798	0.794	0.851	0.849	0.847	0.843	**0.839**	0.737	0.739	0.737	**0.734**	**0.731**
mode	0.727	0.727	0.726	0.727	0.726	0.725	0.732	0.731	0.730	0.729	0.691	0.693	0.692	0.692	0.691
SVD (50/500/50)	0.647	0.648	0.645	0.643	0.636	0.788	0.788	0.788	0.785	0.780	0.456	0.453	0.450	0.449	0.450
missForest	0.662	0.650	0.622	0.593	0.580	0.684	0.673	0.626	0.564	0.521	0.377	0.449	0.442	0.395	0.383

	Colorado Beetle					Arabica Coffee					Wheat				
	(10, 1)	(10, 1)	(5, 1)	(5, 1)	(3, 1)	(15, 3)	(10, 3)	(5, 5)	(3, 10)	(3, 3)	(8, 10)	(5, 10)	(5, 10)	(3, 10)	(3, 10)
ChARF	**0.835**	**0.824**	**0.818**	**0.805**	**0.792**	**0.897**	**0.886**	**0.878**	**0.866**	0.854	0.821	0.808	0.795	0.777	0.762
kNN (50/10/10)	0.765	0.763	0.765	0.765	0.764	0.867	0.866	0.866	0.865	**0.864**	**0.823**	**0.819**	**0.818**	**0.815**	**0.811**
mode	0.761	0.760	0.762	0.761	0.761	0.807	0.804	0.805	0.805	0.804	0.729	0.727	0.729	0.729	0.727
SVD (50/100/200)	0.740	0.737	0.737	0.735	0.734	0.693	0.694	0.696	0.692	0.690	0.622	0.618	0.609	0.600	0.594
missForest	0.352	0.349	0.361	0.326	0.335	0.497	0.480	0.533	0.541	0.586	0.614	0.736	0.746	0.756	0.755

Conclusion

ChARF is an effective method for missing value imputation in SNP data

- Low time complexity $\mathcal{O}(\frac{p}{\Delta} \cdot \Delta) \Longrightarrow$ works for high-dimensional datasets.
- No need for complete data.

ekaterina.antonenko@polytechnique.edu