Uso de autoencoder para comprimir PDFs de redshifts de galáxias

Gabriel Teixeira

Work in progress

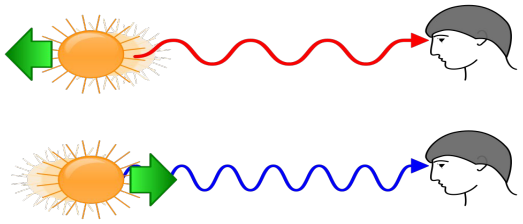
CBPF
Centro Brasileiro de Pesquisas Físicas
UNIDADE DE PESQUISA DO MCTI

Sumário

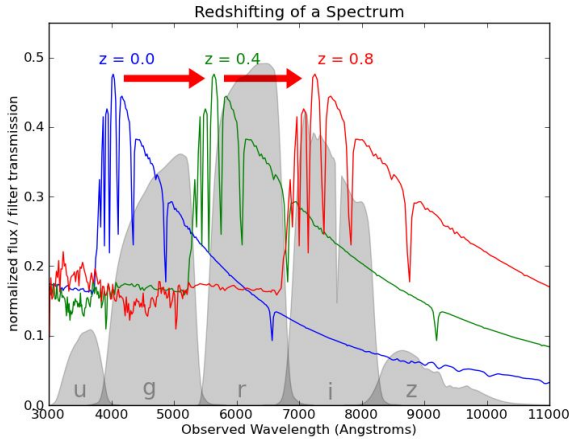
- Introdução - Redshifts
- Obtenção das PDFs
- Problemáticas
- Compactação de PDFs inteiras
- Compactação de geradores
- Ideias futuras

Redshifts

- Redshifts são medidas que mostram o quanto uma galáxia está se afastando de nós.
- Podemos obter o valor de redshift de uma galáxia através do seu espectro luminoso.



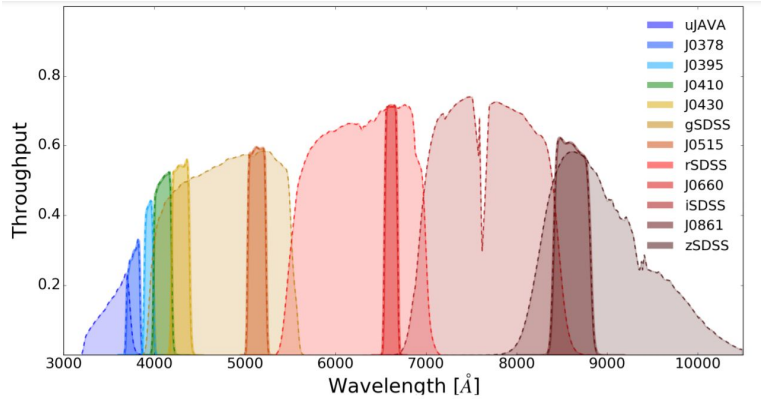
$$z = \frac{\lambda - \lambda_o}{\lambda_o}$$

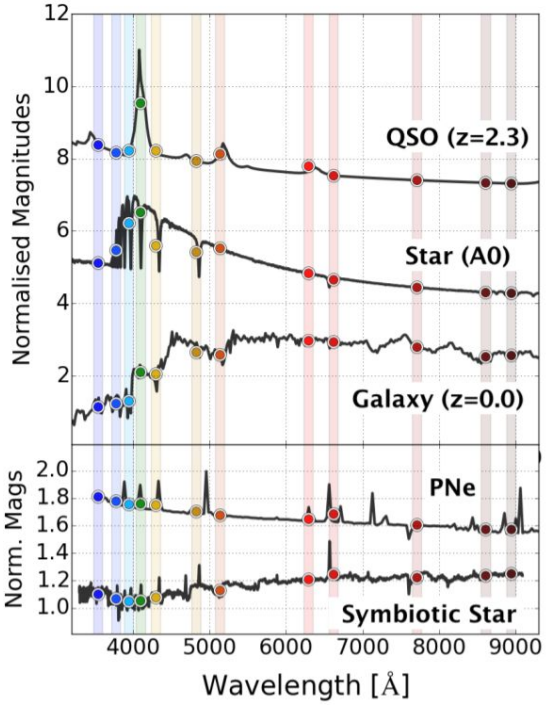


https://ogrisel.github.io/scikit-learn.org/sklearn-tutorial/tutorial/astronomy/regression.html

Redshifts fotométricos

- Diferente da espectroscopia, aqui estamos interessados na magnitude em determinadas bandas do espectro.
- As magnitudes são medidas através de filtros.
- Redshifts fotométricos também são chamados de *photo-z*

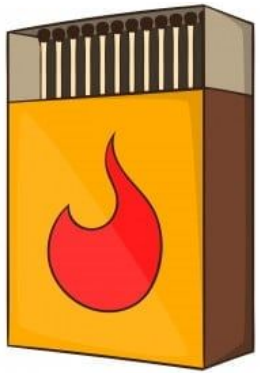


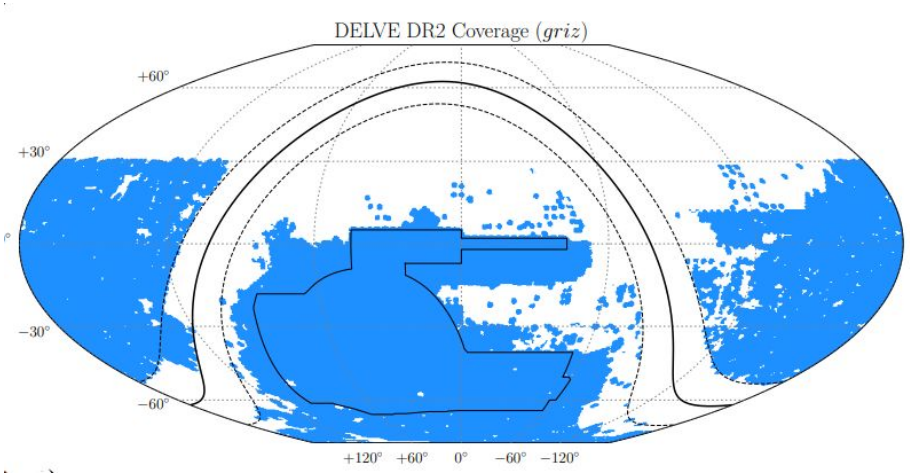


The Southern Photometric Local Universe Survey (S-PLUS): improved SEDs, morphologies and redshifts with 12 optical filters - astro-ph.GA, 2 sep 2019

Como foram obtidos os photo-z's *

- Utilizando dados espectroscópicos e fotométricos do Legacy Survey DR9 (região coberta pelo DELVE Survey DR2)
- Magnitudes das bandas *g, r, z, w1 e w2*



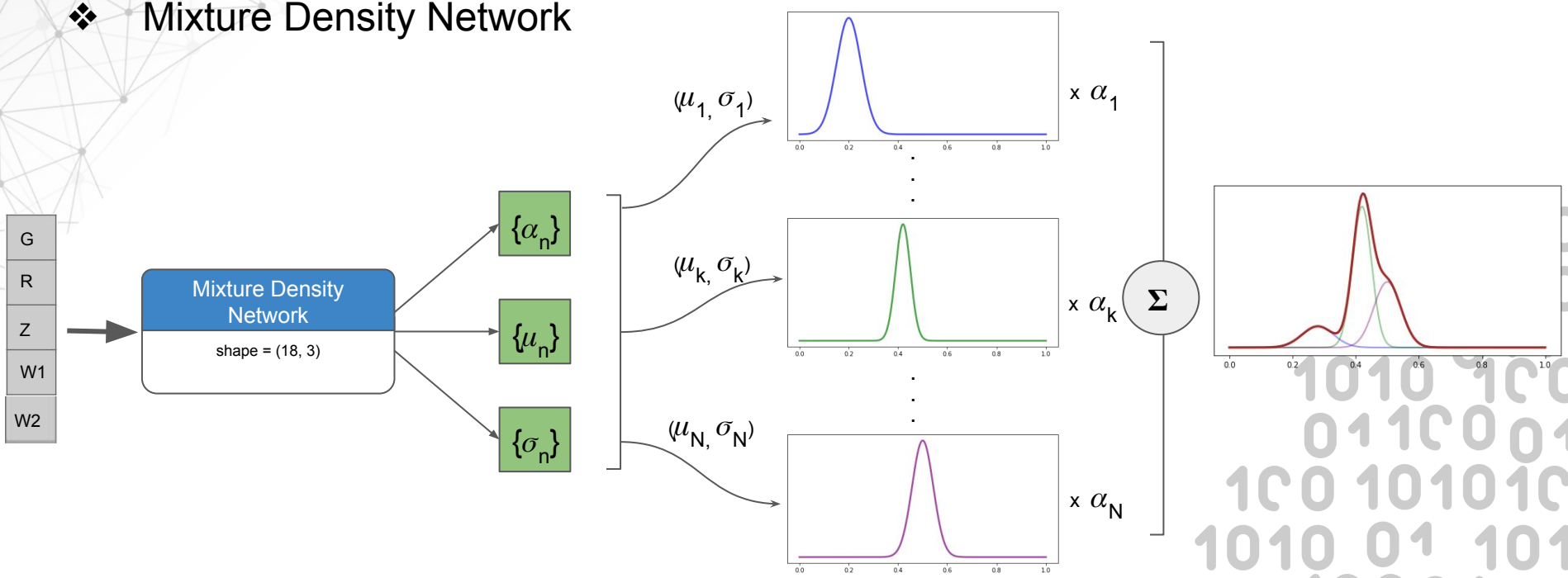


*Para o presente trabalho

https://datalab.noirlab.edu/delve/
https://www.legacysurvey.org/dr9/description/#photometry

Como foram obtidos os photo-z's *

- Mixture Density Network



*Para o presente trabalho

Problemas

- O catálogo fotométrico do DELVE cobre uma área >20.000deg² , possuindo ~17.000deg² com cobertura *g,r,i,z*
- ~10% dos dados disponíveis correspondem a aproximadamente 200,000,000 objetos

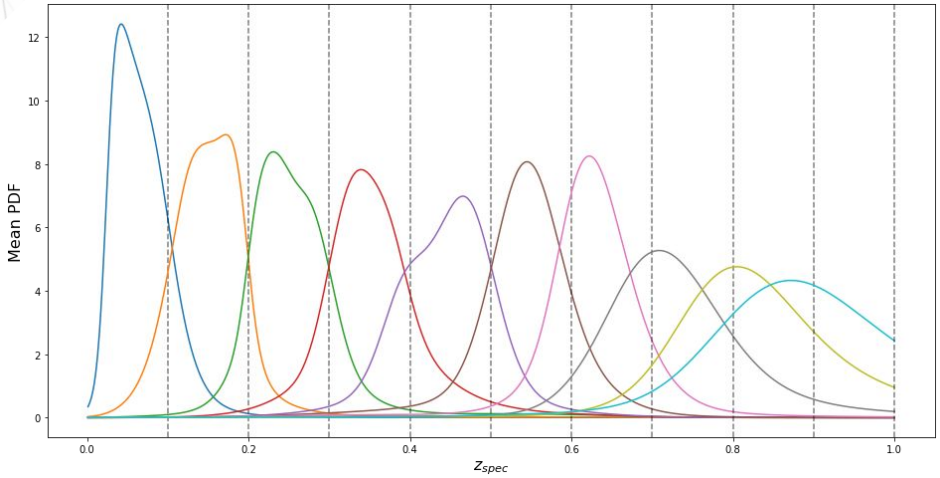
```
>>> fitsfile = Table(train_data['pdf'][:1000, :999])
>>> fitsfile.write('1000PDFs_Table', format='fits', overwrite=True)
```

```
~$ du -hs  1000PDFs_Table
7,8M    1000PDFs_Table
```

- PDFs desses mesmos ~10% então corresponderiam a aproximadamente 1.5 T de armazenamento em disco

Dataset e Pré-processamento

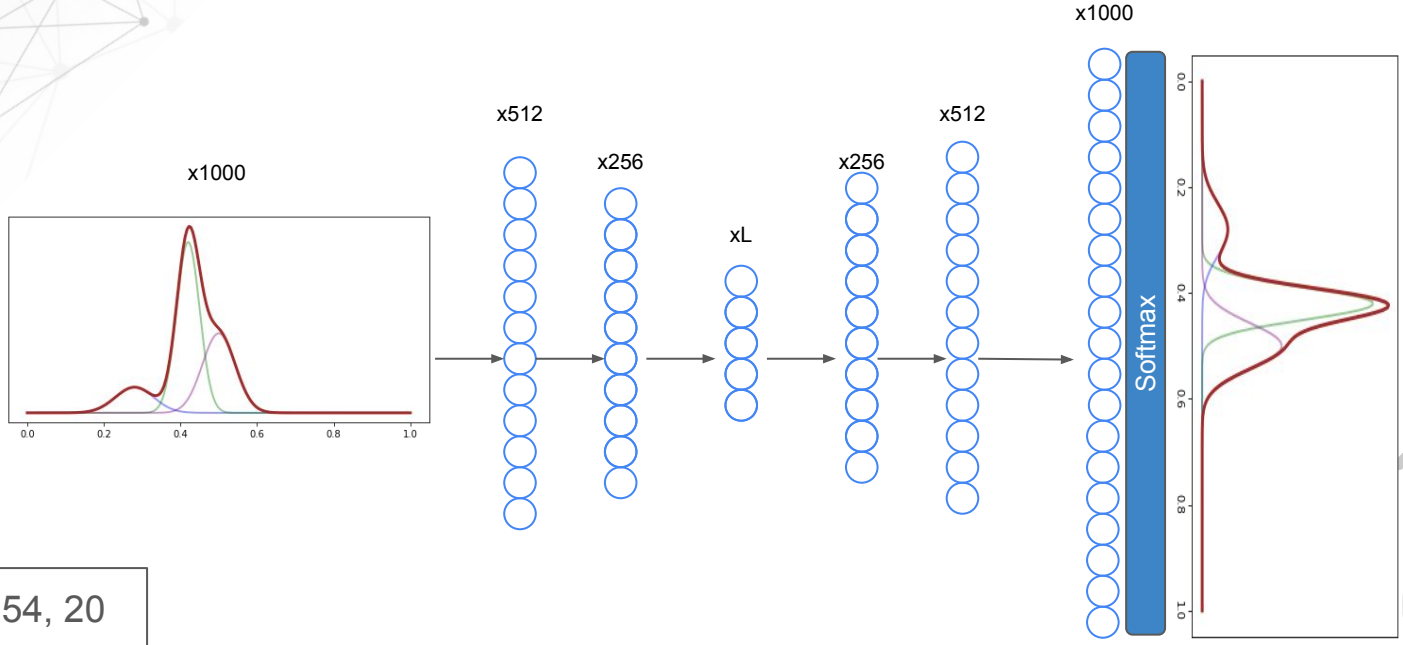
- PDFs geradas pelo modelo MDN
  - 647981 objetos para treino
  - 71862 para validação
  - 503606 para teste



- Input:
  - PDF dividida pela soma de todas as entradas de cada (valores de 0 a 1)

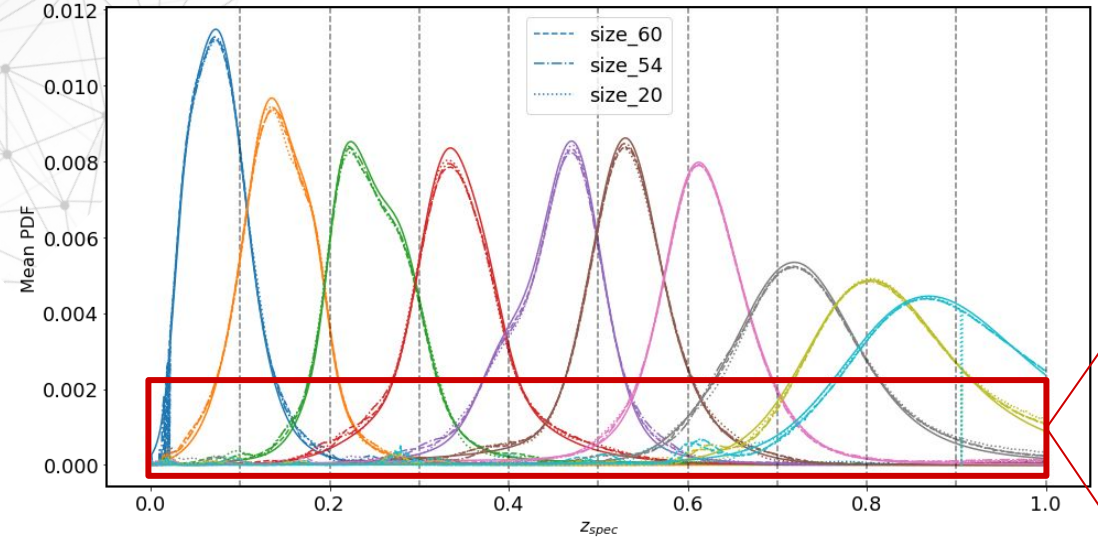
- Output:
  - PDF reconstruída. Saída de uma softmax, valores de 0 a 1

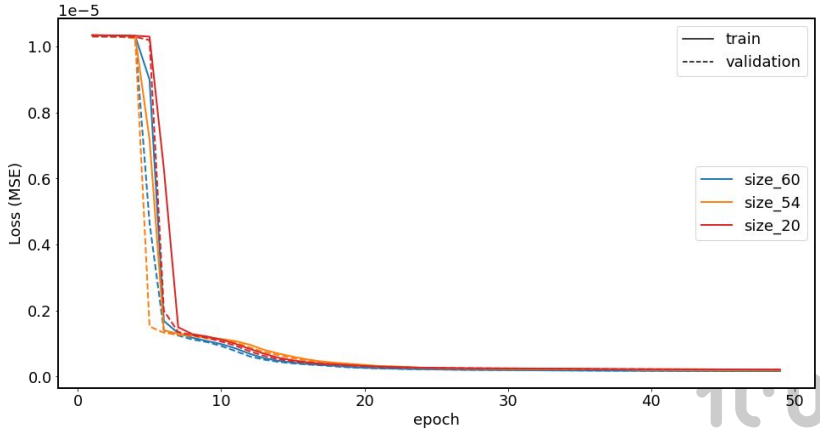
Metodologia - Autoencoder

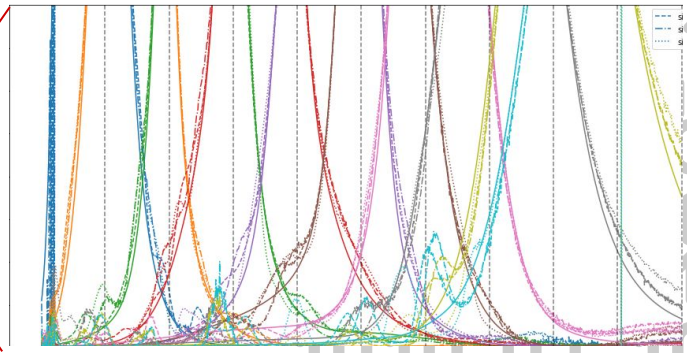


L = 60, 54, 20
Loss -> MSE Épocas -> 50 Batch -> 256

Resultados


size_60
size_54
size_20
Mean PDF
z_{spec}


1e−5
train
validation
size_60
size_54
size_20
Loss (MSE)
epoch



Resultados - Métricas

❖ Point-like Metrics

bias

$$\delta z = z_{phot} - z_{spec}$$

dispersão

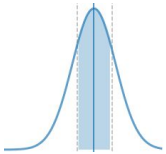
$$\sigma_{NMAD} = 1.48 \times mediana(\frac{|\delta z - mediana(\delta z)|}{1+z_{spec}})$$

median bias

$$median(\delta z)$$

outlier fraction

$$\eta = \frac{|\delta z|}{1+z_{spec}} > 0.15$$



relative error

$$\frac{z_{err}}{1+z_{phot}}$$

❖ PDFs Metrics

PITT

$$PIT = \int_0^{z_{spec}} dz\, PDF(z)$$

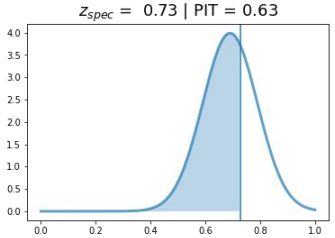
Odds

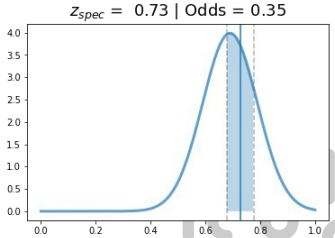
$$Odds = \int_{z_-}^{z_+} dz\, PDF(z)$$

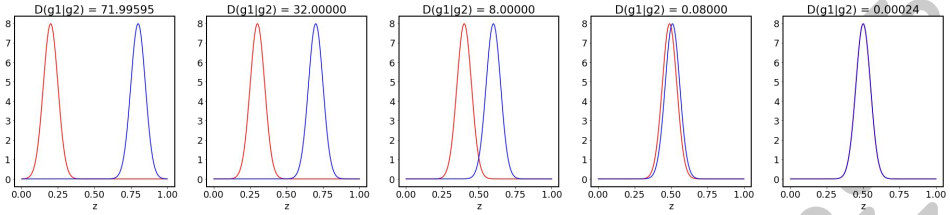
$$z_\pm = z_{spec} \pm 0.06$$

kullback-leibler divergence

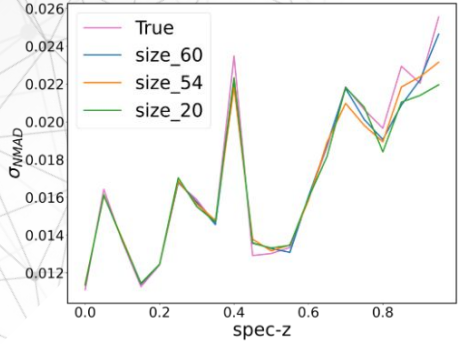
$$D(P_1, P_2) = \int dz P_1(z) log(\frac{P_1(z)}{P_2(z)})$$

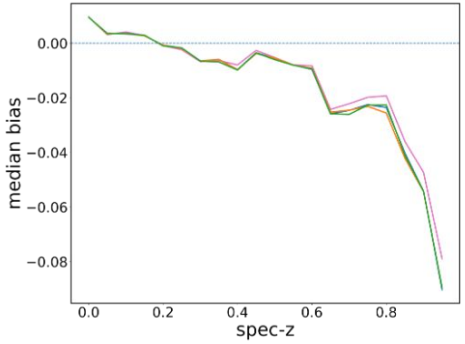


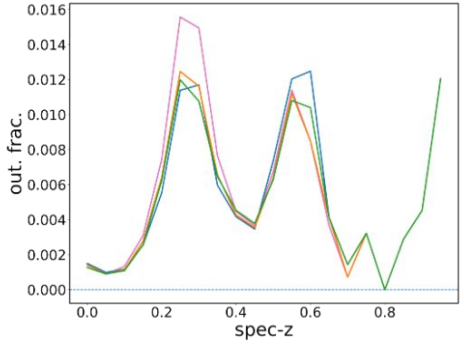


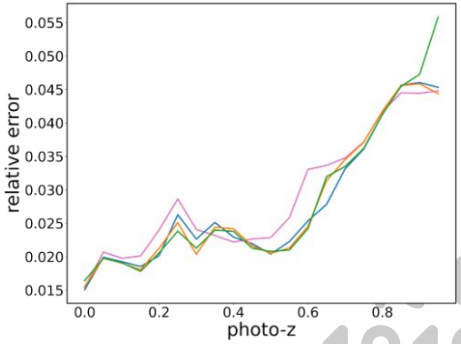


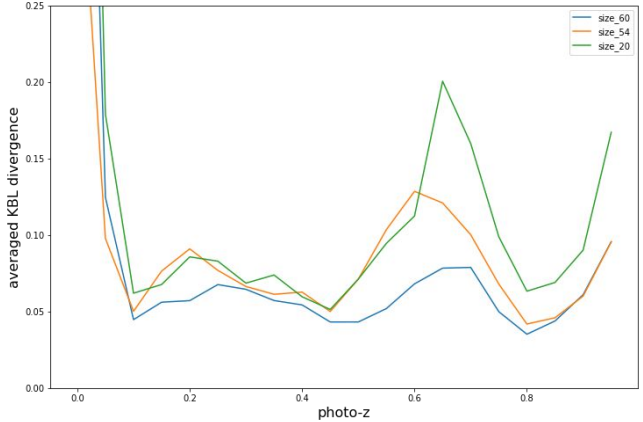
Resultados - Metricas

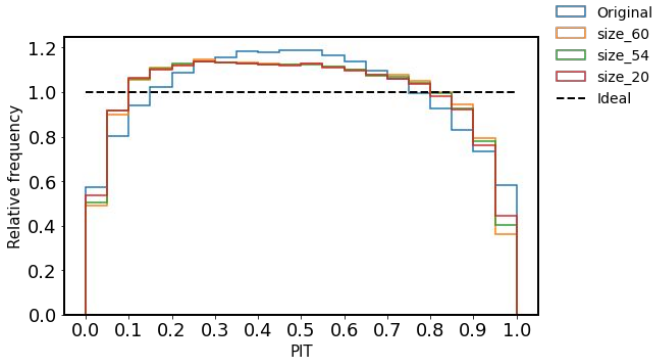


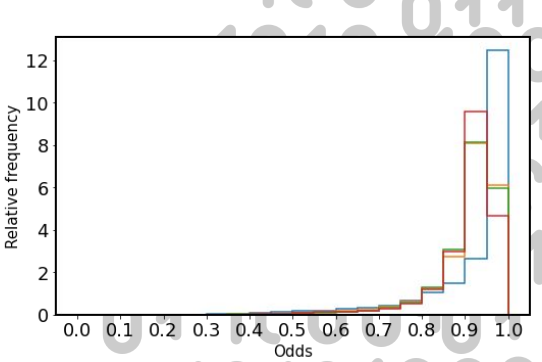












Conclusões - PDF Autoencoder

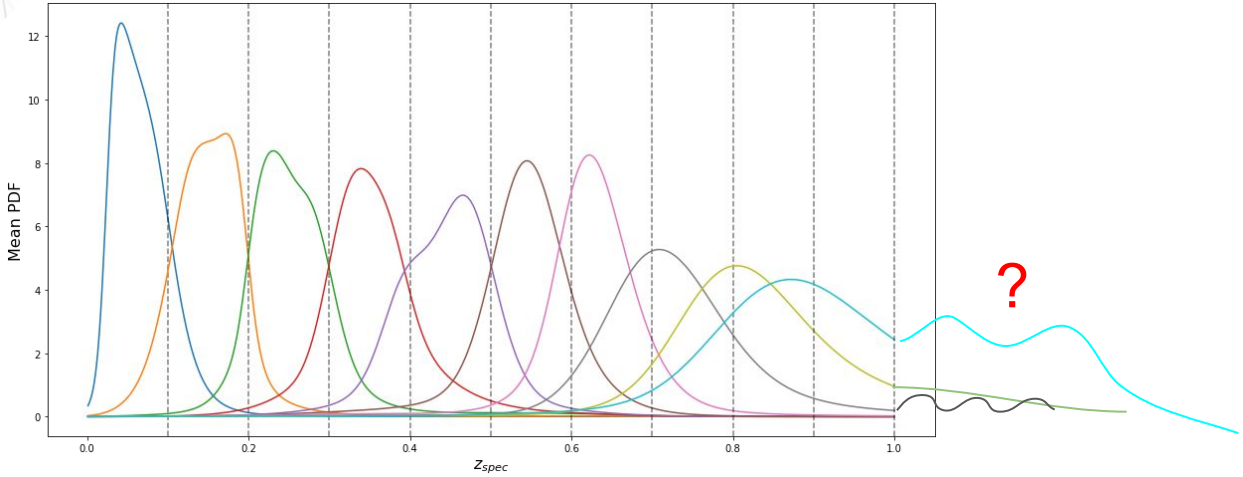
- O método é promissor, porém ainda apresenta divergências relevantes para redshifts mais altos
- As reconstruções das métricas são muito similares entre si para os diferentes valores de L (possível mínimo local?)
- Se utilizássemos a compactação para L=20 o espaço estimado de 1.5 T se reduziria a 30 G

```
~$ du -hs  20Values_Table
168K    20Values_Table
```

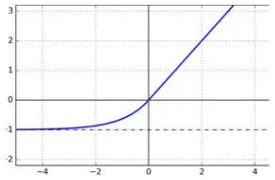
$\{\alpha_n\}$

$\{\mu_n\}$

$\{\sigma_n\}$

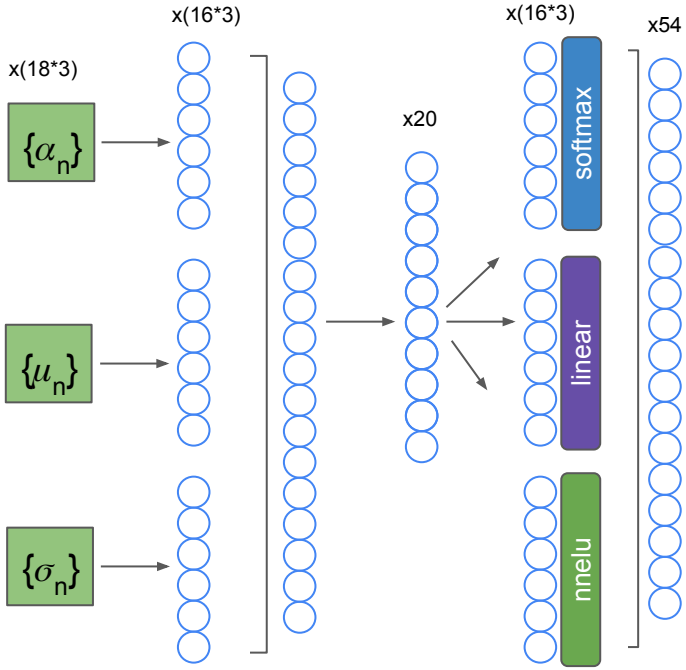


Compactação dos geradores das PDFs



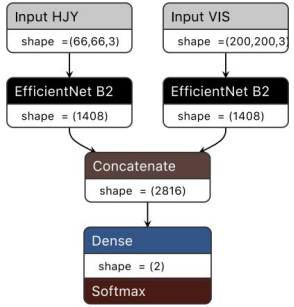
ELU

$$\begin{cases} x & x \geq 0 \\ \alpha(e^x - 1) & x < 0 \end{cases}$$



Loss -> MSE
Épocas -> 30
Batch -> 256

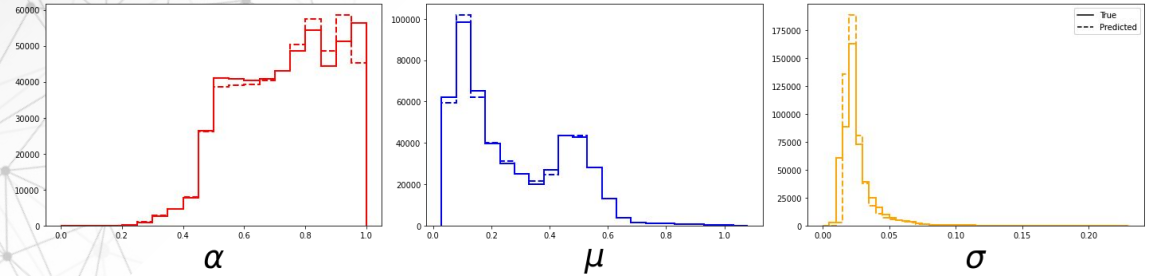
C. R. Bom et al.

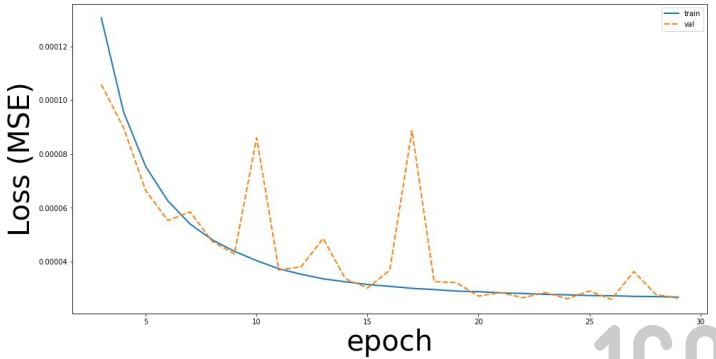


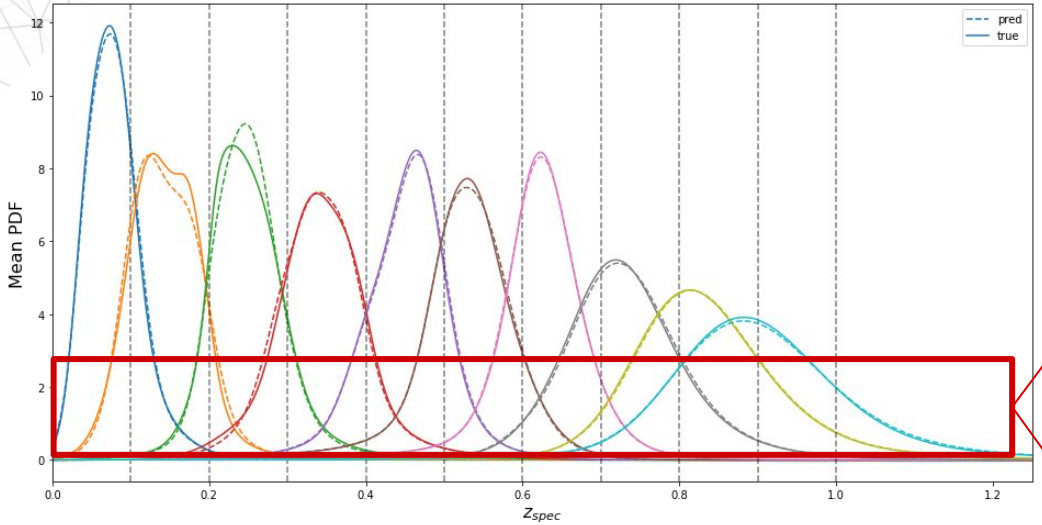
Inspiração

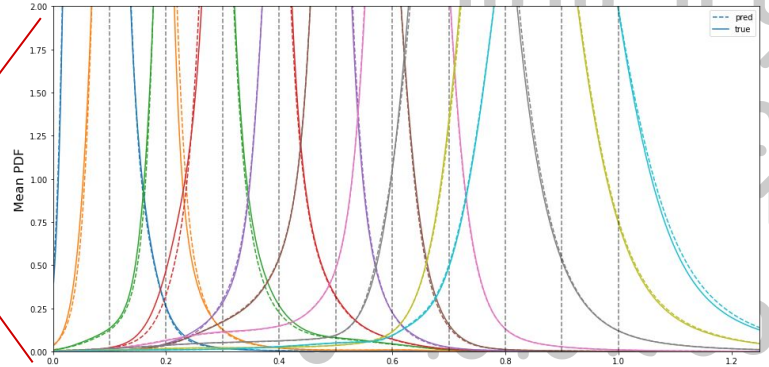
(a)

Resultados

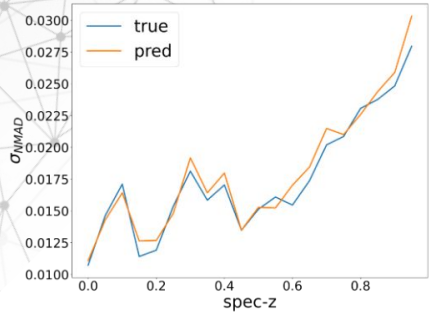

α
μ
σ
True
Predicted

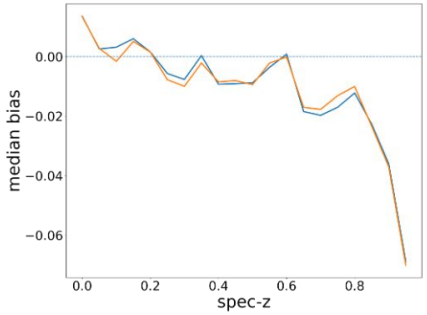

Loss (MSE)
epoch
train
val

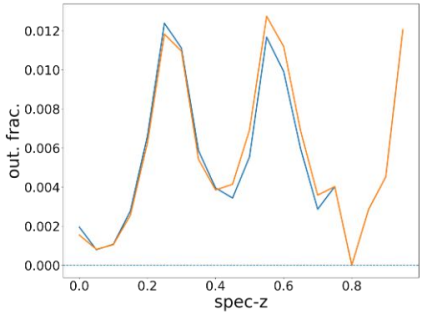

Mean PDF
z_{spec}
pred
true

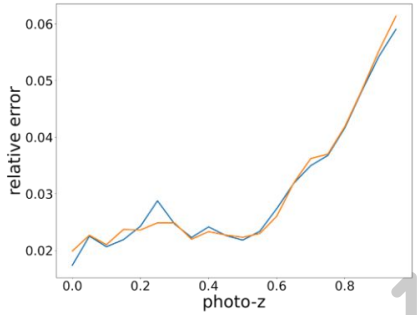

Mean PDF
pred
true

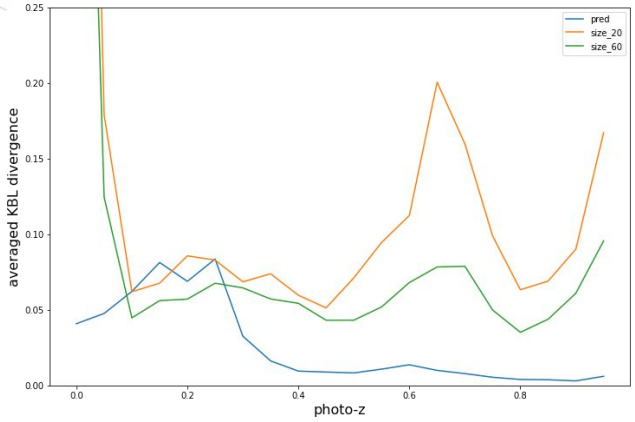
Resultados - Metricas

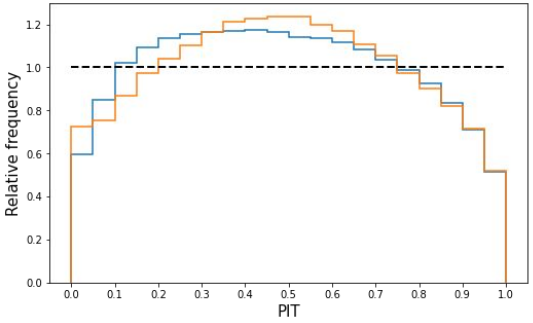


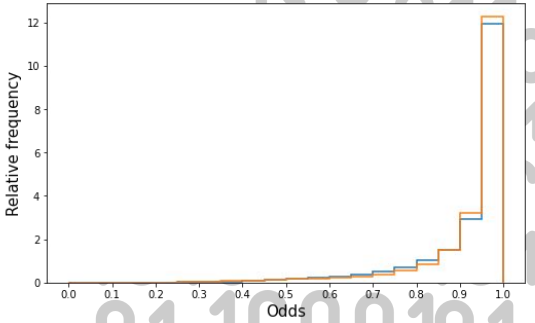












Conclusões - Components Autoencoder

- De fato a ideia de compactar os geradores das PDFs melhora o comportamento das mesmas em alto redshift
- Conseguimos lidar com o problema de precisar de um domínio para definir a PDF
- Obtivemos PDFs mais próximas das originais ao compactar os geradores
- Divergência entre as medidas de PIT ainda precisam ser investigadas mais a fundo

Ideias Futuras

- Utilizar camadas convolucionais (1D) no primeiro método
- Estudar e reforçar a robustez dos métodos com cross validation
- Testar formas de pré-processar os geradores de PDF
- Testar a implementação da divergência KBL como loss function!
- Podemos treinar com muito mais dados !!

Referências

[1] - LIMA, Erik V. R. - Photometric redshifts for S-PLUS using machine learning techniques, 2019.

[2] - Christopher M Bishop. Mixture density networks. 1994.

[3] - https://datalab.noirlab.edu/query.php?name=delve_dr2.photoz

[4] - C R Bom, B M O Fraga et al. , Developing a victorious strategy to the second strong gravitational lensing data challenge, *Monthly Notices of the Royal Astronomical Society*, Volume 515, Issue 4, October 2022, Pages 5121–5134, https://doi.org/10.1093/mnras/stac2047

[5] - Divergência de Kullback-Leibler: Uma aplicação à Modelagem , Jéssica Franco Cançado Richard, Departamento de Estatística

[6] - https://github.com/cdebom/AstroStatistic2021class/tree/main/02_PDF_comp