ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

Positive and Scale Invariant Gaussian Process Latent Variable Model for Astronomical Spectra

Nikolaos Gianniotis, Iliana Isabel Cortés Pérez and Kai Lars Polsterer*

Astroinformatics - HITS gGmbH
Schloss-Wolfsbrunnenweg 35, D-69118 Heidelberg, Germany

Abstract. We propose a probabilistic model that reduces the dimensionality of positive-valued data in a scale-invariant way, treating data items that differ only in scaling as identical. Extending the Gaussian Process Latent Variable Model, we ensure positive function values by applying a non-linear transformation to latent function values. To address the intractable marginal log-likelihood, we utilize a variational lower bound and amortized inference to reduce the number of variational parameters. We apply our model to reconstructing partially observed spectra and show how its scale-invariant property leads to better reconstructions.

1 Introduction

Gaussian processes (GPs) [1] are a versatile tool in machine learning, but imposing on them constraints like positivity, monotonicity, or physical constraints is challenging [2]. Past works have considered constraining GPs as solutions to differential equations [3], temporal and spectral reconstruction problems [4], or injecting domain-specific constraints via linear operators [5]. Other works constrain GP outputs with non-linear functions [6, 7], bound outputs to positive values by constraining the marginal likelihood [8], or cast linear constraints as conditional expectations of the truncated multivariate Gaussian distribution [9].

In this work, we aim to discover a latent space for positive-valued astronomical spectra. Amongst past works on dimensionality reduction for spectra [10, 11, 12], [13] uniquely incorporates a non-negativity constraint. We extend the Gaussian process latent variable model (GPLVM) [14] by bounding its outputs to positive values. The amplitude of astronomical spectra is not an intrincic physical property and should not be reflected in the latent space. We introduce scale-invariance and show that it leads to better reconstructions.

1.1 Gaussian Process latent variable model

We consider the inference of a function $f(\cdot)$ from observed input-output pairs $\{\boldsymbol{x}_n \in \mathbb{R}^Q, y_n \in \mathbb{R}\}_{n=1}^N$. We take a probabilistic view and model $f(\cdot)$ with a GP:

$$f(\boldsymbol{x}) \sim \mathcal{GP}(m, k(\boldsymbol{x}, \boldsymbol{x}'; \boldsymbol{\theta})) \ ,$$

where $m \in \mathbb{R}$ specifies the unknown mean value of $f(\cdot)$ and $k(\boldsymbol{x}, \boldsymbol{x}'; \boldsymbol{\theta})$ specifies the covariance function [1]. A simple regression model for multidimensional

*The authors gratefuly acknowledge the support of the Klaus Tschira Foundation. Software available under `https://github.com/HITS-AIN/GPLVMplus.jl`

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

outputs $\boldsymbol{y}_n \in \mathbb{R}^D$, models each d-th output dimension with an independent GP:

$$\log \int \prod_{d=1}^{D} \prod_{n=1}^{N} \mathcal{N}(y_{nd}|f_{nd}, \beta^{-1}) \mathcal{N}(\boldsymbol{f}_d|\boldsymbol{0}, \boldsymbol{K}) \boldsymbol{df} = \sum_{d=1}^{D} \log \mathcal{N}(\boldsymbol{y}_d|\boldsymbol{0}, \boldsymbol{K} + \beta^{-1}\boldsymbol{I}_N) \ ,$$

where f_{nd} are the function values, and all D dimensions share the same $N \times N$ covariance $\boldsymbol{K}$ (for brevity we suppress the dependence of $\boldsymbol{K}$ on $\boldsymbol{x}$ in the notation). This model can be also employed for dimensionality reduction and is known as the GPLVM [14]: we view $\boldsymbol{y}$ as the images of unknown, latent low-dimensional input coordinates $\boldsymbol{x}$ under a probabilistic map.

1.2 Basic statements

We use the following in the formulation of our model:

$$\langle \ln \mathcal{N}(\boldsymbol{y}|\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}, \boldsymbol{C}) \rangle_{\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})} = \ln \mathcal{N}(\boldsymbol{y}|\boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}, \boldsymbol{C}) - \frac{1}{2} \operatorname{tr}(\boldsymbol{\Sigma}\boldsymbol{A}^T\boldsymbol{C}^{-1}\boldsymbol{A}) \ .$$

We use the expectations wrt $\mathcal{N}(x|\mu, \sigma^2)$, $E(a, \mu, \sigma, b) \equiv \langle \exp(ax+b) \rangle = \exp(a\mu + \frac{(a\sigma)^2}{2} + b)$ and $B(a, \mu, \sigma, b) \equiv \langle \exp(ax + b)^2 \rangle = E(2a\mu, 2a\sigma, 2b)$. The variance of $\exp(ax + b)$ is $V(a, \mu, \sigma, b) \equiv B(a, \mu, \sigma, b) - E(a, \mu, \sigma, b)^2$.

2 Proposed model

We extend the GPLVM to enforce positivity and scale invariance. We call the extension GPLVM$_+$. The map between low-dimensional coordinates $\boldsymbol{x}$ to high-dimensional data $\boldsymbol{y}$ is modelled with D independent GPs. We impose positivity, by applying the $\exp(a \cdot f + b)$ function on the latent function values f of the map, where $a \geq 0$, b respectively control the amplitude and mean value of f. Additionally, we ensure scale invariance by assigning a scaling coefficient $s_n > 0$ to each data item, so that two data items differing only in scale are projected to the same location $\boldsymbol{x}$. The marginal log-likelihood of GPLVM$_+$ reads:

$$\ln \int \prod_{d=1}^{D} \prod_{n=1}^{N} \mathcal{N}(y_{nd}|s_n \exp(a f_{nd} + b), \beta^{-1}) \mathcal{N}(\boldsymbol{f}_d|\boldsymbol{0}, \boldsymbol{K}) \boldsymbol{df}_d \ . \tag{1}$$

2.1 Derivation of lower bound

The introduced non-linearity prohibits analytically integrating out the latent values f_{nd}. We resort to variational inference [15] to approximate the marginal log-likelihood with a lower bound. Denoting the collection of all observations y_{nd} as $\boldsymbol{Y}$, latent values f_{nd} as $\boldsymbol{F}$ and x_{nq} as $\boldsymbol{X}$, the lower bound reads:

$$\ln \int p(\boldsymbol{Y}, \boldsymbol{F}|\boldsymbol{X}) \boldsymbol{dF} \geq \int q(\boldsymbol{F}) \ln p(\boldsymbol{Y}, \boldsymbol{F}|\boldsymbol{X}) \boldsymbol{dF} - \int q(\boldsymbol{F}) \ln q(\boldsymbol{F}) \boldsymbol{dF} \ ,$$

where $q(\cdot)$ is a density we need to postulate, which if equal to the true posterior, makes the lower bound equal to the marginal log-likelihood. We postulate a

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

factorised posterior $q(\boldsymbol{F}) = \prod_{d=1}^{D} q(\boldsymbol{f}_d)$, where $q(\boldsymbol{f}_d) = \mathcal{N}(\boldsymbol{f}_d|\boldsymbol{\mu}_d, \boldsymbol{\Sigma})$, with $\boldsymbol{\mu}_d \in \mathbb{R}^N$ and all D dimensions share the same $N \times N$ covariance $\boldsymbol{\Sigma}$. In order to reduce the number of free parameters in $\boldsymbol{\Sigma}$, we parametrise, according to [16], as $\boldsymbol{\Sigma} = (\boldsymbol{K}^{-1} + \boldsymbol{\Lambda})^{-1}$, where $\boldsymbol{\Lambda}$ is a $N \times N$ diagonal matrix of positive elements. For this choice, the lower bound reads:

$$\langle \sum_{d=1}^{D} \sum_{n=1}^{N} \ln \mathcal{N}(y_{nd}|s_n \exp(af_{nd} + b), \beta^{-1}) \rangle_q + \langle \sum_{d=1}^{D} \ln \mathcal{N}(\boldsymbol{f}_d|\boldsymbol{0}, \boldsymbol{K}) \rangle_q + \mathcal{H}[q] \ , \quad (2)$$

where the entropy term reads $\mathcal{H}[q] = \frac{D}{2} \ln \left((2\pi e)^N |\boldsymbol{\Sigma}| \right)$. Given the basic results in Section 1.2, the first expectation in Eq. (2) reads:

$$\sum_{d=1}^{D} \sum_{n=1}^{N} \frac{1}{2} \ln(\frac{\beta}{2\pi}) - \frac{\beta}{2}(y_{nd} - s_n E(a, \mu_{nd}, \sqrt{\Sigma_{nn}}, b))^2 - \frac{\beta s_n^2}{2} V(a, \mu_{nd}, \sqrt{\Sigma_{nn}}, b) =$$

$$\sum_{d=1}^{D} \sum_{n=1}^{N} \ln \mathcal{N}(y_{nd}|s_n E(a, \mu_{nd}, \sqrt{\Sigma_{nn}}, b), \beta^{-1}) - \frac{\beta s_n^2}{2} V(a, \mu_{nd}, \sqrt{\Sigma_{nn}}, b) \ .$$

The second expectation in Eq. (2) reads $\sum_{d=1}^{D} \ln \mathcal{N}(\boldsymbol{\mu}_d|\boldsymbol{0}, \boldsymbol{K}) - \frac{D}{2} \operatorname{tr}(\boldsymbol{K}^{-1}\boldsymbol{\Sigma})$. To learn the latent space, we optimise the lower bound wrt the free variational parameters $\boldsymbol{\mu}_d$, $\boldsymbol{\Lambda}$, kernel hyperparameters $\boldsymbol{\theta}$, latent coordinates $\boldsymbol{x}_n$ and parameters s_n, a, b, β. We also note that only $\boldsymbol{K}$ depends on $\boldsymbol{x}_n$.

2.2 Amortised inference

The number of free variational parameters in posterior $q(\boldsymbol{F})$ grows with the number of observations: $N \cdot D$ for the means $\boldsymbol{\mu}_d \in \mathbb{R}^N$ and N for $\boldsymbol{\Lambda}$. This leads to a computational burden on the optimizer. We also note that in the lower bound in Eq. (2), the latent coordinates $\boldsymbol{x}$ in matrix $\boldsymbol{K}$ are decoupled from the latent function values f_{nd}, i.e. changing $\boldsymbol{x}$ does not influence directly f_{nd}, which exarcebates the optimisation problem.

We address both issues via amortised inference [17]: we parametrise the variational parameters $\boldsymbol{\mu}_d \in \mathbb{R}^N$ as the images of the latent coordinates $\boldsymbol{x}_n$ under a neural network $g(\boldsymbol{x}_n; \boldsymbol{w}) : \mathbb{R}^Q \to \mathbb{R}^D$, where $\boldsymbol{w}$ are the weights of the network. If g_d is the d-th output of the network, then $g_d(\boldsymbol{x}_n; \boldsymbol{w})$ returns μ_{nd}. Evaluating g_d on all $\boldsymbol{x}_n$, we obtain the vector $\boldsymbol{\mu}_d \in \mathbb{R}^N$. Thus, we optimise the lower bound of Eq. (2) wrt the weights $\boldsymbol{w}$ of network g, $\boldsymbol{\Lambda}$, the latent coordinates $\boldsymbol{x}_n$ and parameters $s_n, a, b, \beta, \boldsymbol{\theta}$ using scaled conjugate gradients[18].

2.3 Inference for test data

Given N_* number of yet unseen, test data items $\boldsymbol{y}_n^*$, we want to infer the corresponding latent coordinates $\boldsymbol{x}_n^*$. The corresponding latent function values are f_{nd}^*. The likelihood for the test data reads:

$$\prod_{n=1}^{N_*} \prod_{d=1}^{D} \mathcal{N}(y_{nd}^*|s_n^* \exp(af_{nd}^* + b), \beta^{-1}) \ . \quad (3)$$

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

The latent function values $\boldsymbol{f}_d^*$ of the test data depend on the latent function values $\boldsymbol{f}_d$ of the training data as follows [1]:

$$p(\boldsymbol{f}_d^*|\boldsymbol{f}_d) = \mathcal{N}(\boldsymbol{f}_d^*|\boldsymbol{K}_*^T\boldsymbol{K}^{-1}\boldsymbol{f}_d, \boldsymbol{K}_{*,*} - \boldsymbol{K}_*^T\boldsymbol{K}^{-1}\boldsymbol{K}_*) \ , \tag{4}$$

where $\boldsymbol{K}_* \in \mathbb{R}^{N_*\times N}$ is the cross-covariance between test and training data; $\boldsymbol{K}_{*,*} \in \mathbb{R}^{N_*\times N_*}$ is the covariance for the test data items[1]. Integrating out $\boldsymbol{f}_d$ in Eq. (4) with respect to posterior $q(\boldsymbol{f}_d)$ reads:

$$p(\boldsymbol{f}_d^*) = \mathcal{N}(\boldsymbol{f}_d^*|\boldsymbol{K}_*^T\boldsymbol{K}^{-1}\boldsymbol{\mu}_d, \boldsymbol{K}_{*,*} - \boldsymbol{K}_*^T(\boldsymbol{K}+\boldsymbol{\Lambda}^{-1})^{-1}\boldsymbol{K}_*) \ .$$

Density $p(\boldsymbol{f}_d^*)$ acts as the prior for the latent function values for the testing data:

$$\ln \int \prod_{n=1}^{N_*}\prod_{d=1}^{D} \mathcal{N}(y_{nd}^*|s_n^*\exp(af_{nd}^*+b), \beta^{-1}) p(\boldsymbol{f}_d^*)\boldsymbol{df}_d \ . \tag{5}$$

We note the similarity of Eq. (5) to the marginal log-likelihood in Eq. (1). Again, we face an intractable integral and we resort to a variational lower bound: we postulate an approximating posterior $r(\boldsymbol{F}^*) = \prod_{d=1}^{D} r(\boldsymbol{f}_d^*)$, $r(\boldsymbol{f}_d^*) = \mathcal{N}(\boldsymbol{f}_d^*|\boldsymbol{\nu}_d, \boldsymbol{A})$. We parametrise the covariance as $\boldsymbol{A} = (\boldsymbol{C}^{-1}+\boldsymbol{L})^{-1}$, where $\boldsymbol{L}$ is a $N_*\times N_*$ diagonal matrix of positive elements [16] . The lower bound reads:

$$\langle \ln \prod_{n=1}^{N_*}\prod_{d=1}^{D} \mathcal{N}(y_{nd}^*|s_n^*\exp(af_{nd}^*+b), \beta^{-1}) + \ln \prod_{d=1}^{D} p(\boldsymbol{f}_d^*)\rangle_r + \mathcal{H}[r] \ , \tag{6}$$

where $\mathcal{H}[r]$ is the entropy of $r(\boldsymbol{F}^*)$. The expectations in Eq. (6) are calculated in the same way as the ones in Section 2.1. To infer latent coordinates $\boldsymbol{x}_n^*$ for the test data $\boldsymbol{y}_n^*$, we optimise the lower bound in Eq. (6) wrt $\boldsymbol{\nu}_d$, $\boldsymbol{L}$, $\boldsymbol{x}_n^*$ and s_n^*.

3 Numerical Experiments

We use a two-hidden layer network $g(\cdot;\boldsymbol{w})$, each layer has 50 units, with tanh activations. We employ the kernel $k(\boldsymbol{x},\boldsymbol{x}';\boldsymbol{\theta}) = \theta_1^2\exp(-\theta_2^2\|\boldsymbol{x}-\boldsymbol{x}'\|^2)$.

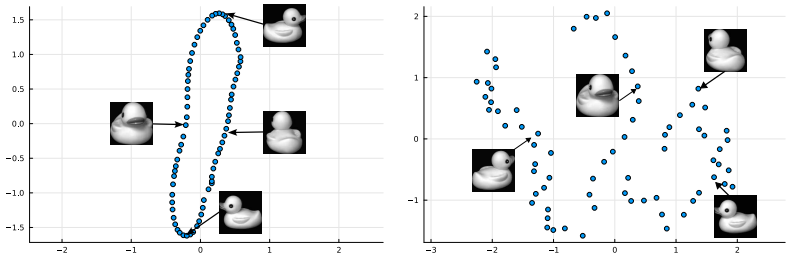
Rubber duck. We use 64×64 images of a rubber duck photographed from 72 angles of a full rotation [19]. The images form an intrinsically two-dimensional closed curve. We scale each image with a factor sampled from the uniform density $\mathcal{U}(0.5, 1.5)$. We train the GPLVM and GPLVM$_+$ on the images for $Q = 2$. Fig. 1 demonstrates the scale invariance of the proposed model.

Spectra. We work with astronomical spectra from the SDSS survey[2] observed on a grid of 500 wavelengths. Each spectrum is observed only in some wavelengths, the rest are treated as missing values. We test the ability of the model to reconstruct partially observed test spectra: in each test spectrum we randomly remove 10% of the values and reserve them for checking model predictions. We

[1] These matrices are actually functions of the latent coordinates i.e. $\boldsymbol{K}_*(\boldsymbol{X}^*)$ and $\boldsymbol{K}_{*,*}(\boldsymbol{X}^*)$. We refrain from denoting them as such in order to lighten notation.

[2] `https://www.sdss.org/`. We shift spectra to the restframe.

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.



(a) GPLVM$_+$ (b) GPLVM

Fig. 1: GPLVM$_+$ finds the intrinsic structure of the data, while GPLVM cannot.

fix $Q = 3$. In a first experiment (Fig. 2a, 2b), we randomly take 1000 spectra from the BOSS catalogue for training, and 500 for testing. Spectra contain negative values that are artefacts due to noise or cosmic rays. The GPLVM and GPLVM$_+$ achieve a normalised mean squared error of 0.61 and 0.42 respectively. A 95% bootstrap confidence interval shows that the difference of ~ 0.19 is statistically significant and lies in $(0.098, 0.362)$. In a second experiment, we take 500 spectra from the Stripe-82 catalogue that exhibit *great variation in amplitude*. The scale invariant GPLVM$_+$ produces good reconstructions, while GPLVM produces unrealistic spectra (Fig. 2c), unlike any spectra in the dataset.

4 Conclusion

We introduced the GPLVM$_+$ for dimensionality reduction of positive-valued data with scale-invariance. Numerical experiments validate its effectiveness. Currently, GPLVM$_+$ incurs high inference costs for latent coordinates of new data. We plan to tackle this in future research.

References

[1] C. E. Rasmussen and C. K. I. Williams. *Gaussian processes for machine learning*. MIT Press, 2006.

[2] L. P. Swiler, M. Gulian, A. L. Frankel, C. Safta, and J. D. Jakeman. A survey of constrained gaussian process regression: Approaches and implementation challenges. *Journal of Machine Learning for Modeling and Computing*, 1(2), 2020.

[3] M. Raissi, P. Perdikaris, and G. Em Karniadakis. Machine learning of linear differential equations using gaussian processes. *Journal of Computational Physics*, 348:683–693, 2017.

[4] Felipe Tobar, Lerko Araya-Hernandez, Pablo Huijse, and Petar M. Djuric. Bayesian reconstruction of fourier pairs. *IEEE Transactions on Signal Processing*, 69:73–87, 2021.

[5] Simo Särkkä. Linear operators and stochastic partial differential equations in gaussian process regression. In *ICANN 2011*, pages 151–158, 2011.

[6] B. S. Jensen, J. B. Nielsen, and J. Larsen. Bounded gaussian process regression. In *IEEE International Workshop on Machine Learning for Signal Processing*, 2013.

[7] Edward Snelson, Zoubin Ghahramani, and Carl Rasmussen. Warped gaussian processes. *Advances in neural information processing systems*, 16, 2003.

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

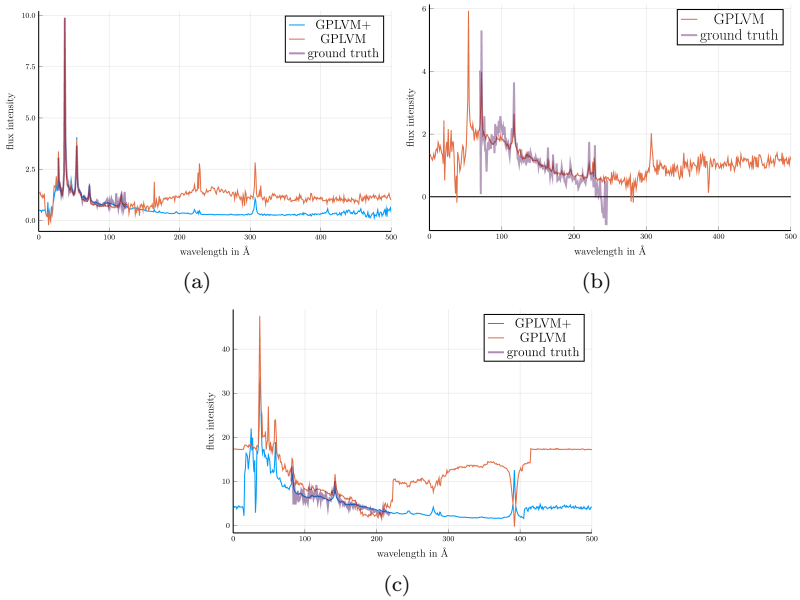


Fig. 2: (a) Both GPLVM and GPLVM$_+$ produce good reconstruction for BOSS data, but (b) GPLVM produces negative values for some cases. (c) GPLVM$_+$ produces good reconstructions for the Stripe-82 data, but the GPLVM does not.

[8] Andrew Pensoneault, Xiu Yang, and Xueyu Zhu. Nonnegativity-enforced gaussian process regression. *Theoretical and Applied Mechanics Letters*, 10(3):182–187, 2020.

[9] S. Da Veiga and A. Marrel. Gaussian process modeling with inequality constraints. *Annales de la Faculté des sciences de Toulouse: Mathématiques*, Ser. 6, 21(3):529–555, 2012.

[10] S. D. Kügler, N. Gianniotis, and K. L. Polsterer. A spectral model for multimodal redshift estimation. In *2016 IEEE SSCI*, pages 1–8, 2016.

[11] S. KN Portillo, J. K Parejko, J. R Vergara, and A. J Connolly. Dimensionality reduction of sdss spectra with variational autoencoders. *The Astronomical Journal*, 160(1):45, 2020.

[12] AC. Eilers, DW. Hogg, B. Schölkopf, D. Foreman-Mackey, FB. Davies, and JT. Schindler. A generative model for quasar spectra. *The Astrophysical Journal*, 938(1):17, 2022.

[13] Markus Harva and Ata Kabán. Variational learning for rectified factor analysis. *Signal Processing*, 87(3):509–527, 2007.

[14] N. Lawrence. Gaussian process latent variable models for visualisation of high dimensional data. *Advances in neural information processing systems*, 16, 2003.

[15] CM. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 2007.

[16] M. Opper and C. Archambeau. The variational gaussian approximation revisited. *Neural Computation*, 21(3), 2009.

[17] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR)*, 2014.

[18] Patrick Kofod Mogensen and Asbjørn Nilsen Riseth. Optim: A mathematical optimization package for Julia. *Journal of Open Source Software*, 3(24):615, 2018.

[19] S. A. Nene, S. K. Nayar, and H. Murase. Columbia object image library (coil-20). Technical report, Technical Report CUCS-005-96, February 1996.