

PhotoZ estimation with Gaussian Processes



Delight in DESC-DC2

Bayesian statistics, Maths and cosmological physics

May 27th 2021

Ref:

Data-driven, Interpretable Photometric Redshifts
Trained on Heterogeneous and Unrepresentative Data
Boris Leistedt^{1,4} and David W. Hogg^{1,2,3}

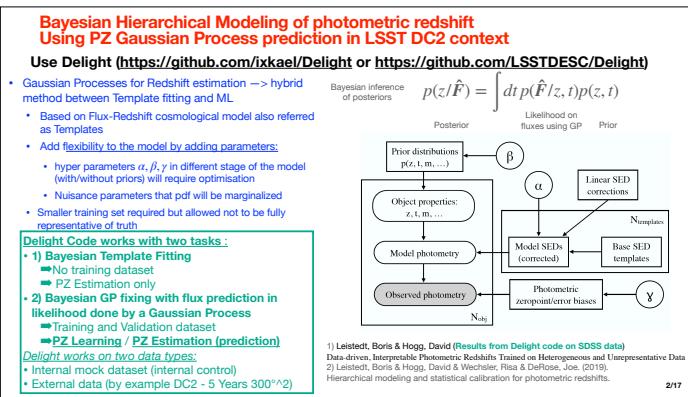
The Astrophysical Journal, 838:5 (14pp), 2017 March 20

<https://doi.org/10.3847/1538-4357/aa6332>

Sylvie Dagoret-Campagne
IJCLab/IN2P3/CNRS

1/17

Je souhaite rapporter sur le travail qui a été fait pour adapter le code d'estimation de PhotoZ Delight au contexte de LSST sur des données DC2.
C'est-à-dire une simulation du survey de LSST de 5 ans dans $300^{\circ}\times 2$.



Dans cette présentation, je présente un code d'estimation PhotoZ proposé par Boris Leistedt et David Hogg exposé dans les articles cités ci-dessous et appelé Delight.

Le principe de ce code repose sur un estimateur Bayesien dit hiérarchique comprenant une fonction de vraisemblance basée sur les flux mesurés, un modèle hiérarchique pour prédire les flux mesurés à partir d'un modèle plus ou moins sophistiqué comme illustré sur le graphique, et l'utilisation de priors sur les distributions de redshift associé à chaque modèle.

Le terme Bayesien hiérarchique signifie qu'on introduit des paramètres à différents niveaux qui apportent de la flexibilité au modèle sous-jacent ou latent flux-redshift qu'on peut classer de la façon suivante:

- Paramètres beta liés aux priors
- Paramètres gamma liés aux biais photométriques, de sélection et d'évolution,
- Paramètres alpha de corrections des modèles de Template SED sous jacents.

Certains paramètres sont des paramètres de nuisance et donc doivent être

marginalisés lors du calcul de la fonction de vraisemblance
 Les autres appelés hyper paramètres doivent être optimisés comme en ML.

Le code Delight comprend deux modes ‘évaluation PhotoZ:

- le Template Fitting qui ne fait que de l’évaluation et donc n’utilise pas de training dataset
- Le Gaussian Process qui évalue le modèle des flux à un redshift donné d’un dataset de validation des galaxies target.

Le GP nécessite au préalable une phase de training et donc un dataset de training.
 Delight propose de fonctionner sur deux types de données:

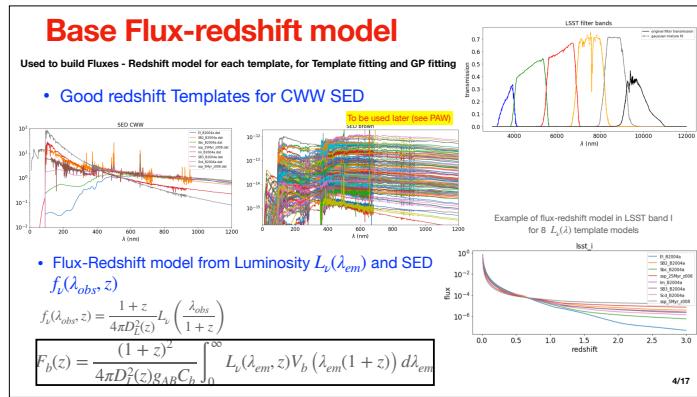
- Des données générées de façon interne de dataset de training et de validation,
- Des données externes. Celui qui est utilisé sont un dataset de DC2.

Bayesian inference of redshift z from noisy fluxes : $\hat{F} = (\hat{F}_1, \dots, \hat{F}_b, \dots, \hat{F}_{N_b})$		
For each target galaxy: $p(z/\hat{F}) = \int dt p(\hat{F}/z, t) p(z, t) \simeq \sum_i \underbrace{p(\hat{F}/z, t_i)}_{\text{Likelihood based on Flux-Redshift model at } z_i \text{ for template } t_i} \underbrace{p(z/t_i)p(t_i)}_{\text{Prior on galaxy template } t_i \text{ and its 2D redshift distribution } p(z, t_i)}$		
Prior	Template Fitting	Gaussian processes
$p(z_i, t_i)$	The redshift priors on SED templates	The redshift priors on redshift taken to be a gaussian at each training galaxy of redshift z_t . See later
Likelihood	$p(\hat{F}/z, t_i)$	Use the analytical Flux-Redshift model $p(\hat{F}/z, t_i) = p(\hat{F}/z, z_t, \hat{F}_t) = \int_{\text{target}} dF p(\hat{F}/F) p(F/z, z_t, \hat{F}_t)$ $p(F/z, z_t, \hat{F}_t) = \mathcal{N}(F - F(z^*); \Sigma_F^*(z))$

Le posterior sur les redshifts à partir des flux mesurés dans les galaxies target s’écrit comme comme une somme du produit de la fonction de vraisemblance pondéré par le prior sur le redshift,

- *Pour le template fitting les priors sont directement liés aux templates de SED. Pour les Gaussian Processes, le prior sont ceux des galaxies du training set.
- * La fonction de vraisemblance est la probabilité de l’erreur sur le flux comparé au modèle.

- Pour template fitting est directement liée à la fonction analytique de relation Flux-Redshift
- Pour le Gaussian Process le modèle est donné par la formule des Gaussian process sur lesquelles nous reviendrons.



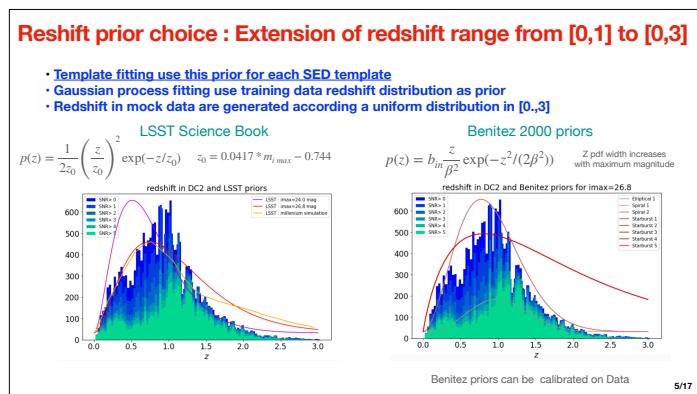
Le modèle flux-redshift de base est construit à partir d'une série de template de SED.

Dans Delight, il y a deux types de SED incluses, les CWW et les Brown.

Pour cette étude, on se limite aux 8 SED CWW pour comprendre comment ça marche.

La formule analytique flux-redshift en fonction de la luminosité et du redshift sont données.

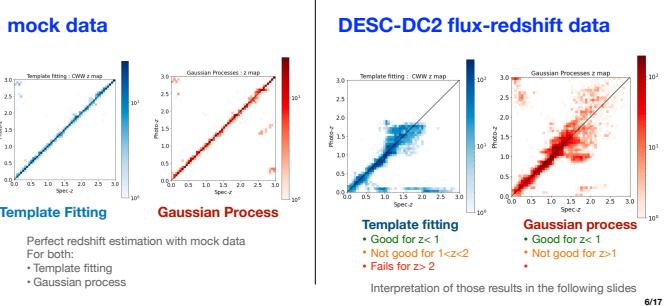
Les transmissions dans les filtres effectifs sont indiqués et la dépendance flux-redshift construite dans la bande i est donnée.



Par rapport aux priors donnés dans Delight il a fallu étendre de domaine de support du redshift pour le rendre compatible avec la distribution dans les données de SED qui dépend de la magnitude limite.

Je donne ici la paramétrisation de la distribution dans le LSST science Book. Mais J'ai choisi la paramétrisation de Benitez pour chacun des 8 SED.

Preliminary PZ results for unoptimized Delight



Voici les résultats du PhotoZ obtenus.

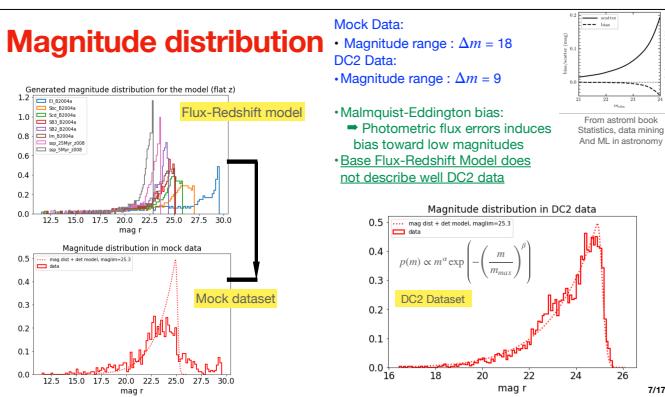
A gauche les résultats pour les mock data générés de façon interne conformément au modèles de SED : le template Fitting et le Gaussian Process fonctionne parfaitement.

Pour les données DC2 à droite, à première vue l'estimation PZ marche moins bien que dans le cas des mock data. Cela est attendu car on n'a pas optimisé les hyperparamètres du code sur le training dataset.

De plus le Template Fitting qui s'écroule au delà de $z>2$ marche moins bien que le Gaussian Process. Cela est probablement dû à la flexibilité du training du Gaussian process alors que le Template Fitting est complètement rigide.

C'est ce qu'on va chercher à comprendre en quoi le GP est meilleurs que le TF dans ce qui suit.

Magnitude distribution



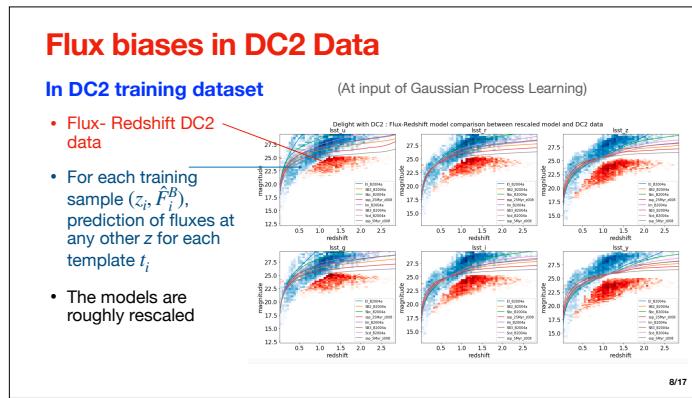
À gauche en haut, voici les magnitudes générées pour les différentes modèles de SED ce qui donne la distribution de magnitudes pour les mock data dans le filtre r très large en magnitude et très différent de celle attendue pour une distribution cosmologique en pointillé.

Au contraire, pour les données DC2, on obtient une distribution plus ramassée conforme à une distribution cosmologique suivant donc le modèle paramètre indiqué ci dessus.

Notons aussi l'effet de biais photométrique proche du seuil de détection le Malmquist-Eddington bias qui en raison des erreurs photométriques, biaise la

galaxies vers les magnitudes élevées en direction des plus faibles ou vers des flux plus élevés.

Ceci nous donne une indication que le modèle flux-redshift de base utilisé n'est pas idéal pour des données de type DC2 et qu'il doit être corrigé.

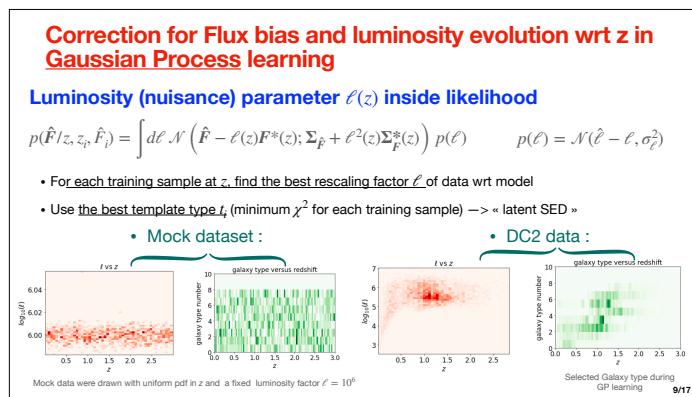


La non conformité du modèle Flux-Redshift de base est confirmée sur ces figures représentant les magnitudes en fonction du redshift dans les différents filtres de LSST et illustre le biais de sélection.

La distribution en rouge correspond aux données DC2.

Les courbes et la distribution en bleue correspondent respectivement au modèle de base Flux redshift et aux valeurs de magnitudes prédictes à chaque redshift pour chacun des modèles de SED.

On voit bien que la place de variation des magnitudes pour les données DC2 est plus restreinte que dans le modèle des SED.



Pour tenir compte de l'évolution du flux avec z ou bien du biais de photo-detection consiste à renormaliser le modèle de base flux-redshift par un paramètre de luminosité ℓ de nuisance.

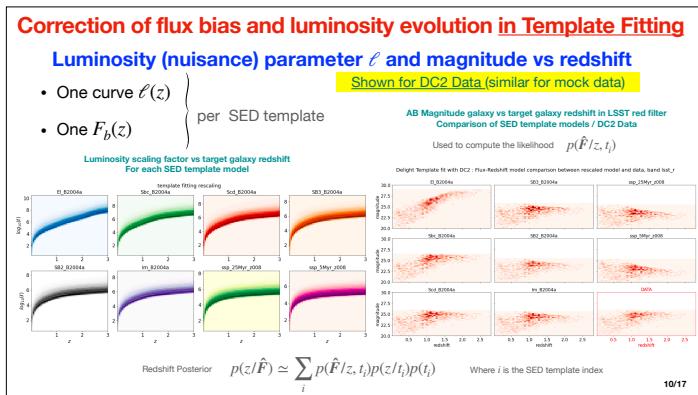
Ce paramètre est appris durant la phase de training du Gaussian Process.

C'est ce que montre les figures à gauche où ℓ est indépendant de z pour le mock dataset comme attendu par construction et à droite pour DC2 pour lesquels ℓ est plus petit à petit z et plus grand à grand z pour tenir compte du biais photométrique.

Cet effet consiste à rehausser le flux dans le modèle à grand z pour compenser le

biais de sélection.

De plus le Gaussian Process sélectionne le Template qui ajuste le mieux les données, pas forcément uniformément pour les données DC2.



Dans le cas du template fitting chaque modèle de template est renormalisé dans la figure de gauche , atténuant ainsi les petits redshifts par rapport au grands redshift. Ainsi dans les figures à droite, on peut comparer la magnitude en fonction du redshift dans les données DC2 dans le cadre rouge par rapport aux modèles renormalisés dans les 8 cadres en noir.

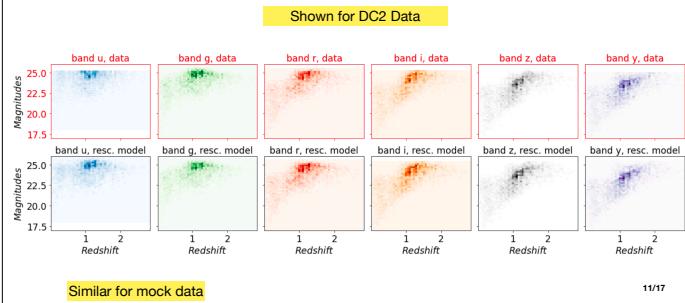
Le posterior est ainsi la somme des likelihoods dont les modèles ont été corrigés en luminosité pondérés par les priors sur des redshifts des modèles.

On s'attend à ce que i soit plus élevé à grand z pour corriger le biais de sélection à grand z .

Imaginons que cette renormalisation soit insuffisante, alors les galaxies dans les données apparaîtraient plus lumineuses que dans le modèle. Le redshift estimé serait donc biaisé vers les petits z .

Correction for bias in Gaussian Process learning

- Comparison of magnitudes/redshift between data and rescaled model



Ici on compare la relation magnitude-redshift pour chaque filtre de LSST établie par la phase d'apprentissage du Gaussian Process.

En haut les données DC2 et en bas le meilleur modèle pour chaque galaxie de training.

Visuellement l'accord données-modèle semble parfait.

Gaussian Process redshift estimation

Compute flux Likelihood on Target galaxy

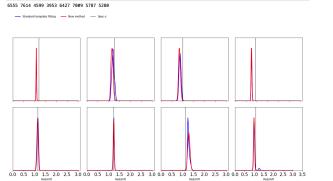
$$p(\hat{F}/z_i, t_i) = p(\hat{F}/z_i, z_i, \hat{F}_i) = \int_{\text{target}} dF p(\hat{F}/F) p(F/z_i, z_i, \hat{F}_i)$$

Using GP prediction

$$p(F/z_i, z_i, \hat{F}_i) = \mathcal{N}(F - F(z)^*, \Sigma_F^*(z))$$

GP posterior on target Galaxy

$$p(z|\hat{F}) = \sum_i p(\hat{F}/z_i, z_i, \hat{F}_i) \stackrel{\text{redshift prior}}{\underset{\text{training}}{\mathcal{N}(z_i, \sigma_i)}} \stackrel{\text{Distribution of the highest evidence over target galaxies}}{\dots}$$



A la phase de trains du Gaussian process, au cours de laquelle un certain nombre de paramètres internes du GP sont sauvegardées pour chacune des galaxies de training, succède la phase estimation.

Les flux de chaque galaxie target supposé au redshift z sont comparés à chacune des galaxies de training considérée tour à tour comme un template au redshift connu z_i .

Le posterior à la fin est la somme des postérieurs sur chacune des galaxies de training.

La figure montre un exemple de posteriors pour le template fitting en bleu et pour

Reminder on what is Gaussian Process and definition of notations

After past introduction of GP by François Fleuret

Find the prediction of the function $y = f(x)$

- for a new value y_* at x_* (n targets)
- from previously m observed training samples (X, y)

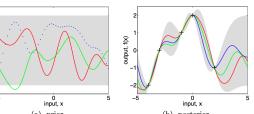
$$p(f_* | X_*, X, y) = \mathcal{N}(\bar{f}_*, \text{cov}(f_*))$$

Standard formula of Gaussian Process for noisy data points (on y)

Average on predicted y_*

$$\bar{f}_* = E[f_* | X, y, X_*] = K(X_*, X)[K(X, X) + \sigma_n^2 I]^{-1}y$$

vector ($n \times 1$)



Learning phase ($m \times m$)
Prediction phase ($n \times m$)
Noise on the m training Data points

The Kernel $K(X, X)$ chosen according to an assumption (or a prior)
Ex: the RBF (Radial Basis Function)

$$k(x_1, x_2) = \sigma^2 \exp\left(-\frac{1}{2l^2}(x_1^T - x_2^T)\right)$$

Covariance on predicted y_*

$$\text{cov}(f_*) = K(X_*, X_*) - K(X_*, X)[K(X, X) + \sigma_n^2 I]^{-1}K(X, X_*)$$

Prediction phase ($n \times n$)

Prediction phase ($n \times m$)

Learning phase ($m \times m$)

Prediction phase ($m \times n$)

13/17

Voici un très bref rappel générique sur les Gaussian Processes.

On les utilise quand on veut trouver une fonction $y=f(x)$ non paramétrique, pour laquelle on dispose de m échantillons de y et d'un X qui peut être multi-dimensionnel. Le GP permet de prédire les propriétés statistiques de y^* en tout x^* selon les formules suivantes.

Ces prédictions dépendent de matrices de Kernel apprises pendant une phase d'apprentissage et certaines évaluées pendant la phase de prédiction.

La forme fonctionnelle de ce Kernel dépend de bonnes raisons apriori liées telles que des hypothèses mathématiques de régularité mais aussi de très bonnes

raisons physiques comme dans notre cas.

Estimation of target redshift with Gaussian Process in Delight

Find a non parametric function $y = f(x)$, where
• Y is the vector of LSST fluxes $F(b, z)$ in the 6 LSST filters,
• X is a complicated vector of a band index b_j , redshift z and luminosity scaling factor ℓ^k for each training & target galaxy.

Training noisy fluxes: $\hat{F} = (\hat{F}_1, \dots, \hat{F}_b, \dots, \hat{F}_{N_b})$, size $(B \times 1)$ and covariance matrix $\Sigma_{\hat{F}}$

Predicted noiseless fluxes: $F^* = (F_1^*, \dots, F_b^*, \dots, F_{N_b}^*)$, size $(B^* \times 1)$ and covariance matrix Σ_F^*

Prior on noiseless model fluxes: $p(F|X) = \mathcal{N}(\mu^F(X), k^F(X, X))$

Standard formula of Gaussian processes for prediction

- Average: $F^* = \mu^F(X^*) + k^F(X^*, X)[k^F(X, X) + \Sigma_F]^{-1} \times (\hat{F} - \mu^F(X))$
- Covariance: $\Sigma_F^* = k^F(X^*, X^*) - k^F(X^*, X)[k^F(X, X) + \Sigma_F]^{-1}k^F(X, X^*)$

The only term available for template fitting Additional term for GP

But what are the chosen expression for μ^F and k^F ?

14/17

From which cosmological concepts μ^F and k^F are derived ?

Luminosity is a linear combination of Template + adding eventual emission lines

Luminosity: $L_b(\lambda, \alpha, I) = \ell^I \sum_t^N \alpha_t T_b^t(\lambda) + \ell^R \frac{R_b(\lambda)}{\text{residuals}}$

SED templates

Residuals: $R_b \sim \mathcal{GP}(0, k^R(\lambda, \lambda'))$

$k^R(\lambda, \lambda') \text{ chosen to be a RBF}$

$L_b(\lambda, \alpha, I) \sim \mathcal{GP}\left(\ell^I \sum_t^N \alpha_t T_b^t(\lambda), \ell^R k^R(\lambda, \lambda')\right)$

Flux: $F_b(z, \alpha, \ell) \sim \mathcal{GP}(\mu^F(b, z, \alpha), k^F(b, b', z, z', \ell, \ell'))$

$\mu^F(b, z, \ell, \alpha) = \frac{\ell(1+z)^2}{4\pi D_L^2(z)g_{AB}} \sum_t^N \int_0^\infty T_b^t(\lambda_{em}, z) V_b(\lambda_{em}(1+z)) d\lambda_{em} = \ell \sum_t^N \alpha_t F_b^t(z)$

$k^F(b, b', z, z', \ell, \ell') = \left(\frac{(1+z)(1+z')}{4\pi D_L(z)D_L(z')g_{AB}}\right)^2 \frac{\ell\ell'}{C_b C_{b'}} \int_0^\infty V_b((1+z)\lambda) V_{b'}((1+z')\lambda') k^R(\lambda, \lambda') d\lambda d\lambda'$

15/17

Que sont les Y et les X dans le cadre de Delight ?

- Les Y ce sont les flux bruités dans les 6 bandes tant pour les training set que target set.
- Les X sont les 6 indices des bandes, le redshift et le facteur de luminosité. X est donc un vecteur multidimensionnel comprenant 6 indices discrets et de 2 variables continues.

Mais ce qui est intéressant c'est la façon dont le Kernel est choisi ou construit.

Le Kernel de Delight est construit de la façon suivante:

- On considère que la luminosité est une somme pondérée de templates de SED à laquelle on rajoute des résidus correctifs, ensemble de continuums et de raies d'émission.
- Ces résidus sont par définition un GP de valeur moyenne nulle et de kernel RBF dans l'espace des longueurs d'ondes.
- On en déduit les muF et kF comme des fonctions d'indice de bande, de z et I et de paramètre de pondération de modèle selon les formules indiquées sur ce slide.

- KF lui même dépend de l'intégration du kernel K-lambda aux longueurs d'onde d'émission multiplié par les fonction de transfert des filtres aux longueurs d'onde de détection.
- Je ne peux pas détailler, mais cette modélisation comprend de nombreux hyperparamètres qui ne demandent qu'à être optimisés sur les données DC2.

Conclusion on this work

- Delight provides a new way for PZ estimation based on GP in the context of Bayesian statistics.
 - ➡ Compromise between ultra flexible ML without priors on physics requiring a very representative training set and rigid Template fitting with «hard » coded physics model in it,
- Extended physical hierarchical model with a moderate number of hyperparameters (understandable physically) requiring a limited training dataset not necessarily fully representative
- Delight standard configuration (for SDSS) has been extended for LSST
 - ➡ Redshift priors extended to redshift [0-3] (used for Template Fitting only)
- Delight works well (Template fit & GP) with mock data (no luminosity evolution and flux bias)
- Delight works not that well for DC2 fluxes by now
 - ➡ Was expected for Template Fitting,
- Results for GP are better than Template fitting but far from optimal however encouraging,
 - ★ Namely No optimization has been performed

16/17

Delight provides a new way for PZ estimation based on GP in the context of bayesian statistics.

Compromise between ultra flexible ML without priors on physics requiring a very representative training set and rigid Template fitting with «hard » coded physics in it,

Extended physical hierarchical model with a moderate number of hyperparameters (understandable physically) requiring a limited training dataset not necessarily representative

Delight standard config (for SDSS) has been extended for LSST

Redshift priors extended to redshift [0-3] (used for Template Fitting only)

Delight works well (Template fit & GP) with mock data (no luminosity evolution and flux bias)

Delight works not that well for DC2 fluxes by now

Was expected for Template Fitting,

Results for GP are better than Template fitting but far from optimal however encouraging,

Namely No optimization has been performed Namely No optimization has been performed

Conclusion / Next steps

- [Optimize GP hyper parameters over DC2 data using CWW SED latent SED](#)
- Extend [SED CWW set to Brown SED](#) and try to optimize again.
- Many path to explore ways [to refine the GP model](#)
 - ➡ Add more emission lines,
 - ➡ Find Other features
 - ➡ [Many new idea for models see Leistedt, Boris & Hogg\(2019\) not implemented in Delight](#)

Try to optimize GP hyper parameters over DC2 data using CWW SED latent SED
Extend SED set to Brown SED and try to optimize
Many path to explore ways to refine the GP model
Add more emission lines,
Find Other features
Many new idea for models see Leistedt, Boris & Hogg(2019) not implemented in Delight