Presentation

A Bayesian approach for regression in the presence of covariate shift: an application to galaxies redshift estimation

Presented by Hieu Cao

Supervisor Prof. Simon Wilson

School of Computer Science & Statistics
Trinity College Dublin

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Outline

Outline

Covariate shift

refering to the situation where the distributions of the independent variables in the training and testing data are different, while the conditional distributions of the dependent variables given the independent ones remain the same between the two sets.

$$p_{train}(x) \neq p_{test}(x),$$
$$p_{train}(y \mid x) = p_{test}(y \mid x)$$

$\Longrightarrow$ How can use training data efficiently to predict in testing set accurately?

Redshift estimation

- Redshift: measure of the increase in wavelength of photons.
- redshift is typically estimated through spectroscopic surveys, which are time- and cost-intensive.
- photometric surveys, which observe more galaxies in a shorter time at a lower cost, by measuring the magnitudes of the light from galaxies through different colored filters.

However, the estimated redshift from photometric surveys is subject to the covariate shift problem (Figure 1).

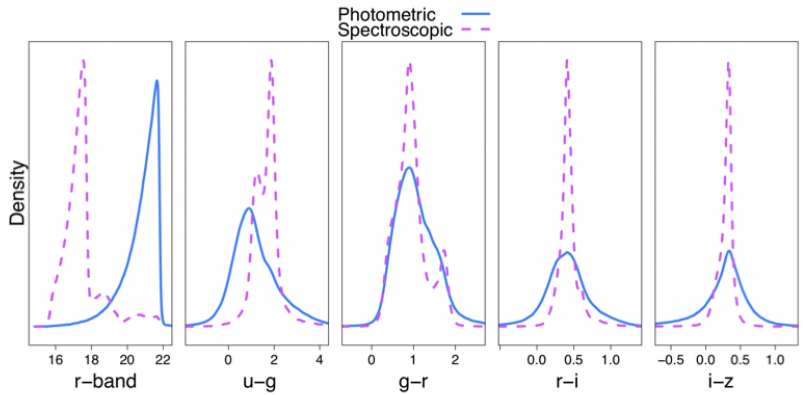


Figure: Covariate shift in cosmology data.[1]

[1] Rafael Izbicki, Ann B. Lee, and Peter E. Freeman. "Photo-z estimation: An example of nonparamet conditional density estimation under selection bias". In: *The Annals of Applied Statistics* 11.2 (2017), pp. 698–724. DOI: 10.1214/16-AOAS1013. URL: https://doi.org/10.1214/16-AOAS1013.

Outline

Hierarchical Bayesian regression model, Sampling method, Variogram estimation

- We adopt Gaussian Processes to model the redshift estimation problem, building on the work of Almosallam et al.[2].
- Gaussian Process :

$$f \sim \mathcal{GP}(m, K)$$

  With $m(\mathbf{x}) = \mathbb{E}(f(\mathbf{x}))$ and $K(\mathbf{x_i}, \mathbf{x_j}) = \mathbb{E}((f(\mathbf{x_i}) - m(\mathbf{x_i}))(f(\mathbf{x_j}) - m(\mathbf{x_j})))$.
- Since the modelling for mean and covariance function is complicated, we use Hamilton Monte Carlo to sample to the posterior.
- Variogram estimation: This method measures the correlation between samples based on the spatial information (additional features) empirically.

[2] Ibrahim A. Almosallam, Matt J. Jarvis, and Stephen J. Roberts. " GPz: non-stationary sparse Gauss processes for heteroscedastic uncertainty estimation in photometric redshifts". In: *Monthly Notices of th Royal Astronomical Society* 462.1 (July 2016), pp. 726–739. ISSN: 0035-8711. DOI: 10.1093/mnras/stw eprint: https://academic.oup.com/mnras/article-pdf/462/1/726/18470400/stw1618.pdf. URL: https://doi.org/10.1093/mnras/stw1618.

Coping with covariate shift

Methods used to cope with covariate shift in our work:

- Kernel mean matching (**KMM**).
- Optimal transport (**OT**).
- Neighbourhood Component analysis (**NCA**).

Kernel mean matching (KMM)

Estimating the ratio $r(x) = \frac{p_{test}(x)}{p_{train}(x)}$
Using moment matching between two distributions : $p_{test}(x)$ and $r(x).p_{train}(x)$. Estimate $r*$ is the MSE optimal solution for :

$$\underset{r}{\text{argmin}} \| \int K(\mathbf{x}, .).r(x)p_{train}(x)dx - \int K(\mathbf{x}, .)p_{test}(x)dx \|^2$$

Since the expectation is computed empirically, we can estimate vector $\mathbf{r}^* := (r^*(x_1), r^*(x_2), \ldots, r^*(x_n))^T$ by $\widehat{\mathbf{r}} = \underset{\mathbf{r}}{\text{argmin}}(\frac{1}{n_p^2}\mathbf{r}^T\mathbf{K}_{\mathbf{tr},\mathbf{tr}}\mathbf{r} - \frac{2}{n_p n_q}\mathbf{r}^T\mathbf{K}_{\mathbf{tr},\mathbf{te}}\mathbb{1})$

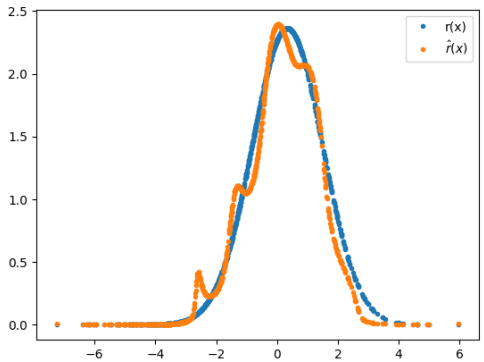


Figure: estimate r of two distributions.

optimal transport (OT)

Monge problem: find the transport map that moves sand mass A (with histogram a) to sand mass B (with histogram b) at a minimum cost (measured as the Euclidean distance between the two masses).
The output of the algorithm is the optimal transport plan $\hat{P}$, which satisfies the marginal constraints $\sum_i \hat{P}_{i,j} = a$ and $\sum_j \hat{P}_{i,j} = b$.

Neighbourhood Component analysis (NCA)

learn a distance function like Mahalanobis distance $d_M(\boldsymbol{x}, \boldsymbol{x}') = \sqrt{(\boldsymbol{x} - \boldsymbol{x}')^T \boldsymbol{M} (\boldsymbol{x} - \boldsymbol{x}')}$ using information from data.
the probability of x_i being a neighbour of x_j is:

$$p_{ij} = \frac{\exp\left(-\|\boldsymbol{L}\boldsymbol{x}_i - \boldsymbol{L}\boldsymbol{x}_j\|_2^2\right)}{\sum_{l \neq i} \exp\left(-\|\boldsymbol{L}\boldsymbol{x}_i - \boldsymbol{L}\boldsymbol{x}_l\|_2^2\right)}, p_{ii} = 0,$$

where $M = L^T L$. And, the probability that x_i correctly classified is: $p_i = \sum_{j:y_j = y_i} p_{ij}$.
The metric is learnt by maximizing $\sum p_i$.

Outline

Generative model

$$z = f(x) \tag{1}$$

$$f \sim \mathcal{GP}(m, K) \tag{2}$$

$$\mathbf{f} = \begin{pmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_n) \end{pmatrix} \sim \mathcal{MVN}\left(\begin{pmatrix} m(x_1) \\ m(x_2) \\ \vdots \\ m(x_n) \end{pmatrix}, \begin{pmatrix} K(s_1, s_1), \ldots, K(s_1, s_n) \\ K(s_2, s_1), \ldots, K(s_2, s_n) \\ \ddots \\ K(s_n, s_1), \ldots, K(s_n, s_n) \end{pmatrix} \right) \tag{3}$$

$$\mathbf{x_{train}} = (x_1, x_2, .., x_t) \sim \mathcal{N}(\mu_0, \sigma_0^2) \tag{4}$$

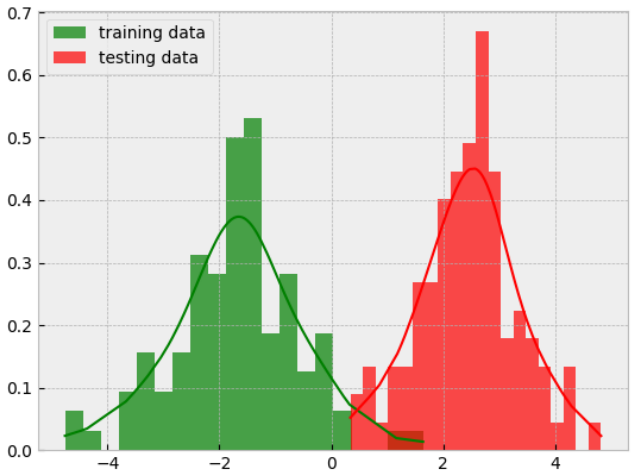
$$\mathbf{x_{test}} = (x_{t+1}, x_{t+2}, .., x_n) \sim \mathcal{N}(\mu_1, \sigma_1^2) \tag{5}$$

$$\text{Mean function: } m(x) = \phi(x)^T \theta, \text{ with } \phi_i(x) = \exp\left\{ -\frac{\|x - p_i\|^2}{\gamma_i^2} \right\} \tag{6}$$

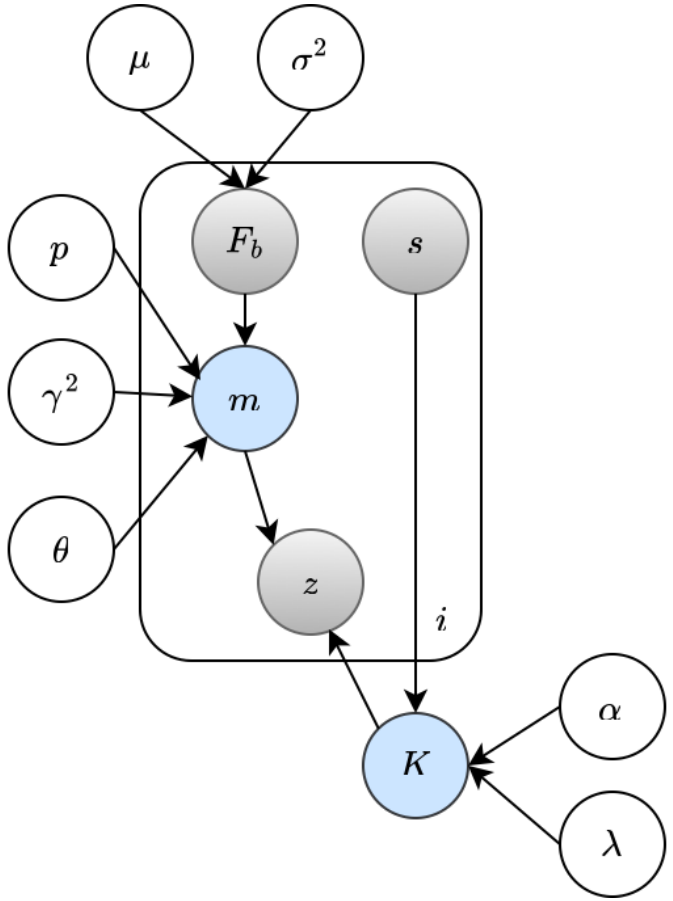
$$\text{Covariance function: } \mathbf{K_{i,j}} = K(\mathbf{s_i}, \mathbf{s_j}) = \lambda \exp\left\{ -\alpha \|\mathbf{s_i} - \mathbf{s_j}\|_2 \right\} \text{ with } \alpha, \lambda > 0 \tag{7}$$

$$\text{with } \|\mathbf{s}_i - \mathbf{s}_j\|_2 = \sqrt[2]{(x_{1i} - x_{1j})^2 + \ldots + (x_{pi} - x_{pj})^2}$$

Generative model



(a) training and testing covariates



(b) Directed acyclic graphical model.

Predictive distribution

$$\text{Posterior distribution: } \mathbf{f}_{test} \mid \mathbf{f}_{train}, \mathbf{x}_{train}, \mathbf{x}_{test}, \mathbf{s} \sim \mathcal{MVN}(\bar{\mathbf{m}}, \bar{\mathbf{K}}) \tag{8}$$

$$\text{With } \bar{\mathbf{m}} = \mathbf{m}(x_{test}) + \mathbf{K}_{te,tr}\mathbf{K}_{tr,tr}^{-1}(\mathbf{f}_{train} - \mathbf{m}(x_{train})) \tag{9}$$

$$\bar{\mathbf{K}} = \mathbf{K}_{te,te} - \mathbf{K}_{te,tr}\mathbf{K}_{tr,tr}^{-1}\mathbf{K}_{tr,te} \tag{10}$$

Since we use the noise-free observations model (Eq 1), the predictive posterior distribution is the same as the posterior distribution above.
Final prediction: take the integral of predictive posterior distribution through the posterior of parameters imperially.

Variance adaptation to cope with covariate shift

Motivated from the *reweighted* loss in Machine learning:

$$\begin{aligned} E_{p_{test}(x)}(l(h(x),y)) &= \sum p_{test}(x) l(h(x),y) dx \\ &= \sum \frac{p_{test}(x)}{p_{train}(x)} p_{train}(x) l(h(x),y) dx \\ &= E_{p_{train}(x)} w l(h(x),y) \end{aligned} \tag{11}$$

Adjusting the likelihood term in Bayesian framework through the variance: it is more certain about the training points near the test set, and the uncertainty is increasing where the training data is far from test data.
Two steps:

- Estimating the reciprocal of of the weight η.
- Adjusting the variance.

Estimating the reciprocal of the weight η

- Kernel mean matching: the output is the vector of ratio of test distribution to train distribution, $\eta = \frac{1}{\widehat{\mathbf{r}}}$
- Optimal transport: $\eta = (\mathbf{C} \odot \widehat{\mathbf{P}})\mathbb{1}$
- NCA: learning a distance metric between labeled and unlabeled data based on assigning pseudo-labels to the data. $d(x_i, \text{test set}) = \sum_{x_j \in \text{test set}} d(x_i, x_j)$.
  $\eta = \mathbf{d}$

Adjusting the variance

- The reciprocal of the weight $\hat{\eta}$ can be transformed by taking min max normalization.

$$\hat{\eta}_i = 1 + \frac{2\eta_i - max(\eta) - min(\eta)}{2(max(\eta) - min(\eta))}$$

- Normal distribution: $\widehat{\sigma}^2 = \sigma^2 \hat{\eta}$
- Multivariate normal distribution:

$$\begin{aligned} \hat{\mathbf{Cov}} &= diag(\sigma \odot \widehat{\eta}^{\circ \frac{1}{2}}).Corr.diag(\sigma \odot \widehat{\eta}^{\circ \frac{1}{2}}) \\ &= (\widehat{\eta}^{\circ \frac{1}{2}})(\widehat{\eta}^{\circ \frac{1}{2}})^T \odot \mathbf{Cov} \end{aligned}$$

Outline

Simulated data

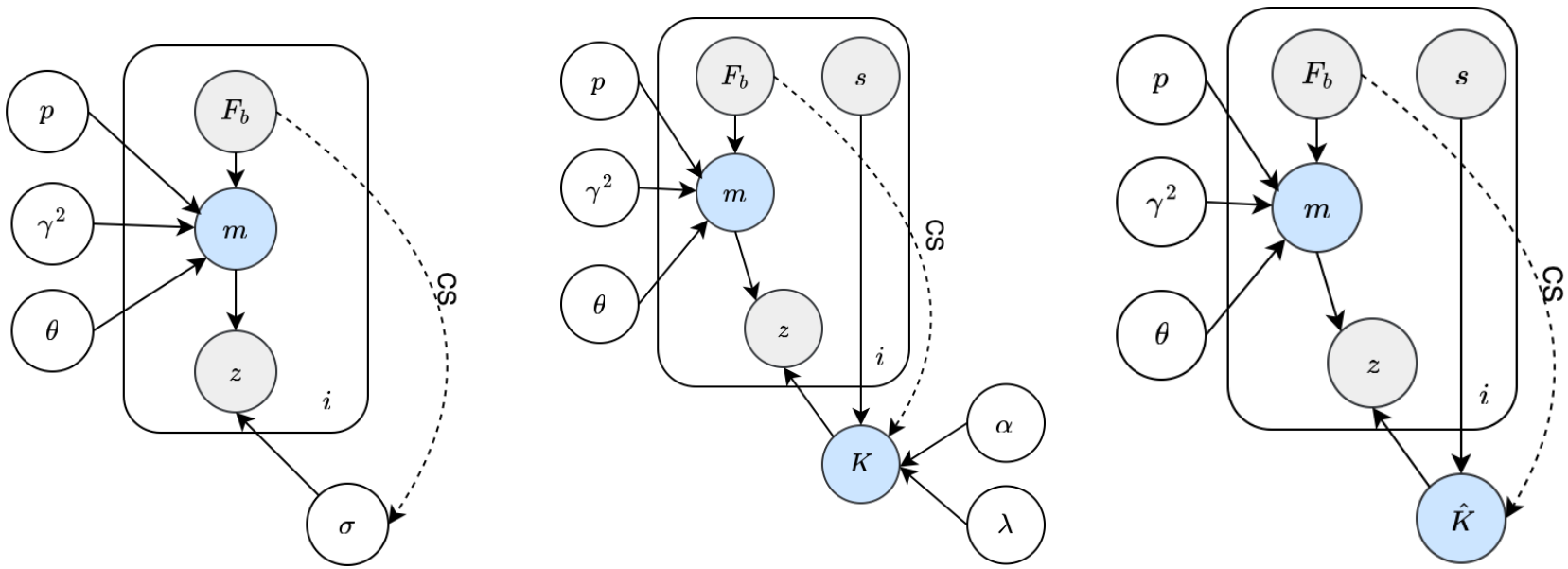
- Additional features (spatial information): generated from Mixture Gaussian.
  $p(s) = \frac{1}{2}\mathcal{N}\left(\begin{pmatrix}-1\\-1\end{pmatrix}, \begin{pmatrix}4 & 1\\1. & 4.\end{pmatrix}\right) + \frac{1}{2}\mathcal{N}\left(\begin{pmatrix}4\\4\end{pmatrix}, \begin{pmatrix}1 & 0.\\0. & 1.\end{pmatrix}\right)$
- Assign these features randomly for 100 training and 100 testing data points.
- Covariates and dependent variables generated as the model in last section.

Simulated data

Table: Parameters for the true generative model.

Parameters	value	Prior
μ_0	-1.6	
μ_1	2.5	
σ_0^2	1.	
σ_1^2	2.	
θ_0	3	$\mathcal{N}(0,100)$
θ_1	-2	$\mathcal{N}(0,100)$
p_0	-1	$\mathcal{N}(0,100)$
p_1	2	$\mathcal{N}(0,100)$
γ_0	1.2	$HalfNormal(0,100)$
γ_1	4.	$HalfNormal(0,100)$
α	0.3	$HalfNormal(0,100)$
λ	0.5	$HalfNormal(0,100)$

Proposed Models



(a) BRM **(b)** BRM + SI **(c)** BRM + SI (Vario)

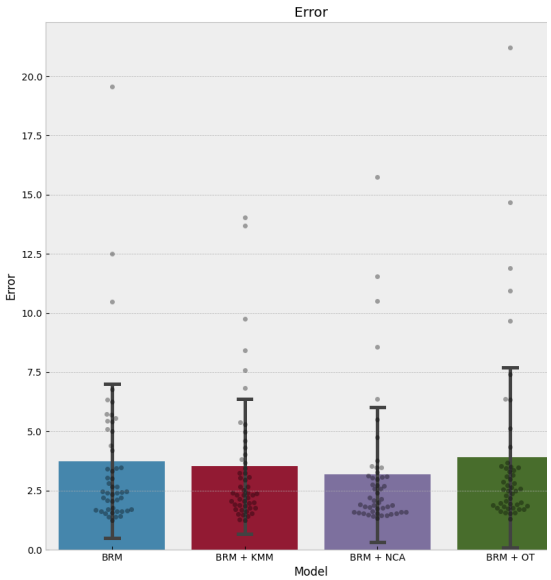
Figure: Graphical models used in experiments. The dashed line and dashed circle indicate the addition parts for covariate shift methods.

Result

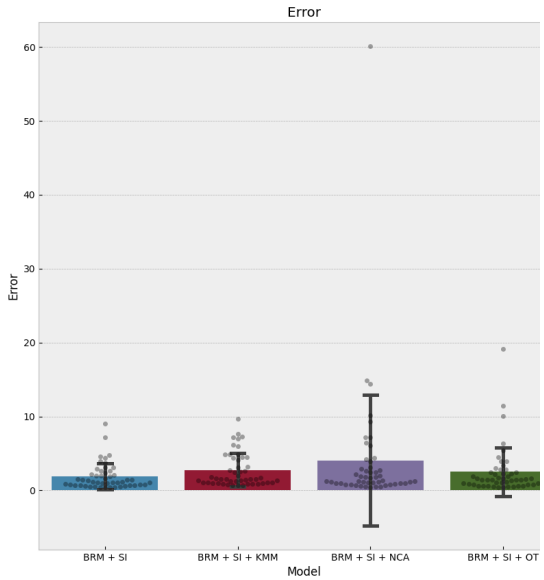
Table: Comparison between different models by mean and standard deviation of MSE through 50 running experiments

Model	Mean	Std
BRM	3.744	3.265
BRM + KMM	3.521	2.855
BRM + NCA	3.179	2.851
BRM + OT	3.903	3.809
BRM + SI	1.871	1.769
BRM + SI + KMM	2.739	2.283
BRM + SI + NCA	4.053	8.838
BRM + SI + OT	2.497	3.292
BRM + SI (Vario)	4.425	2.533
BRM + SI + KMM (Vario)	5.437	2.654
BRM + SI + NCA (Vario)	5.795	4.121
BRM + SI + OT (Vario)	5.559	3.349

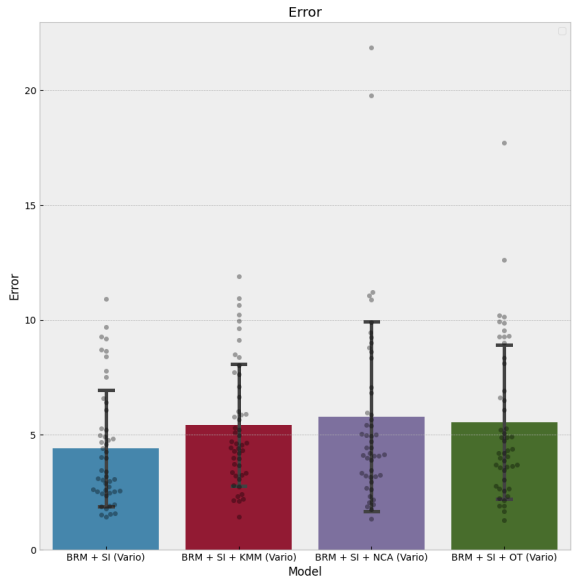
Result



(a) BRM error



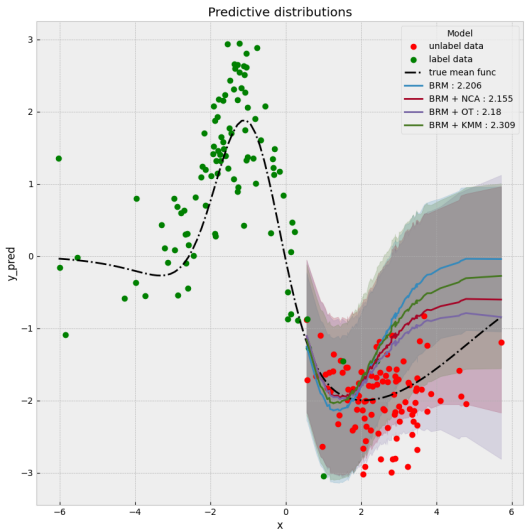
(b) BRM + SI error



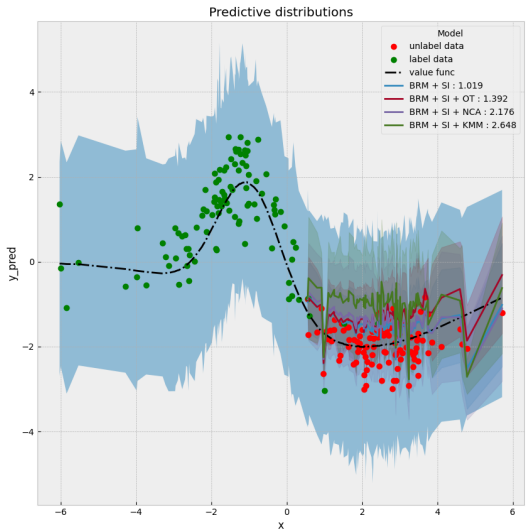
(c) BRM + SI (Vario) error

Figure: The bar chart for MSE of each model, collected through 50 experiments.

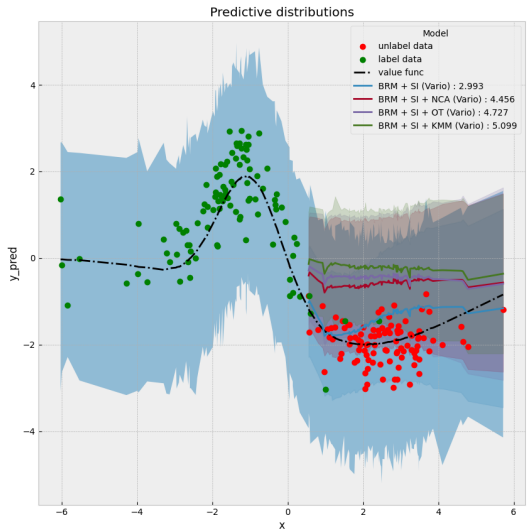
Result



(a) BRM predictions



(b) BRM + SI predictions



(c) BRM + SI (Vario)

Figure: Predictive distributions of models, extracted from one experiment.

Outline

Conclusion

- The good result in applying the KMM and NCA methods to the Bayesian regression model supported for the argument that covariate shift methods only work in case the model is misspecified.
- The variogram estimation error has strong effect to the predictive distribution.
- While the model incorporating spatial information undoubtedly achieved the best results, it's unlikely that we can achieve a truly generative model in this particular case.

Future work

- Exploring the extent of model misspecification required for the covariate shift methods to improve predictions.
- application of the covariate shift methods to cosmological data, which will require expanding models to handle multivariate features.
- Improve model scalability for larger datasets.

Thank you for your listening