Simple, Effective, and General Regression and its Meta-analysis using Kernel Averages

Theory, Challenges, Solutions and Results

Aidan John Mullins, Prof. Chris Oates, Dr. Markus Rau
E-Mail Address: a.j.mullins2@newcastle.ac.uk

Newcastle University

School of Mathematics, Statistics and Physics

Conference Start : 17 September, 2025

Newcastle University

The Problem Setting

$$x \in \mathcal{X} \subseteq \mathbb{R}^{d_{\mathcal{X}}} \longrightarrow \boxed{Q(y|x)} \longrightarrow y \in \mathcal{Y} \subseteq \mathbb{R}^{d_{\mathcal{Y}}}$$

$$\exists\, \theta_0 \in \Theta \;:\; Q = P_{\theta_0} \in \mathcal{P}_\Theta$$

$$\mathcal{P}_\Theta = \{P_\theta \in \mathcal{P}(\mathcal{Y}|\mathcal{X})\}_{\theta \in \Theta}$$

Newcastle University

An Example

$$x \in \mathcal{X} \subseteq \mathbb{R}^{d_{\mathcal{X}}} \longrightarrow \boxed{Q(y|x)} \longrightarrow y \in \mathcal{Y} \subseteq \mathbb{Z}_{\geq 0}$$

$$\exists\, \theta_0 \in \Theta \subseteq \mathbb{R}^{d_{\mathcal{X}}} \;:\; Q = P_{\theta_0} \in \mathcal{P}_{\Theta}$$

$$\mathcal{P}_{\Theta} = \left\{ P_{\theta}\left(y \in \mathcal{Y} \mid x \in \mathcal{X}\right) = \exp\left(-k\left(\theta^{\top} x\right)\right) \sum_{i \in \mathcal{Y}} \frac{k\left(\theta^{\top} x\right)^{i}}{i!} \right\}_{\theta \in \Theta}$$

Newcastle University

Another Example

$$x \in \mathcal{X} \subseteq \mathbb{R}^{d_{\mathcal{X}}} \longrightarrow \boxed{Q(y|x)} \longrightarrow y \in \mathcal{Y} \subseteq \mathbb{R}^{d_{\mathcal{Y}}}$$

$$\exists\, \theta_0 \in \Theta \;:\; Q = P_{\theta_0} \in \mathcal{P}_\Theta$$

$$\mathcal{P}_\Theta = \left\{ P_\theta\,(y \in \mathcal{Y} \mid x \in \mathcal{X}) = \frac{\int_{\mathcal{Y}} \exp\!\left(-\frac{1}{2}(y-m_\theta(x))^\top S_\theta^{-1}(x)(y-m_\theta(x))\right)\lambda(dy)}{\sqrt{2\pi \det(S_\theta(x))}^{d_{\mathcal{Y}}}} \right\}_{\theta\in\Theta}$$

Newcastle University

Existing Methods

- Highly–General Techniques
  - Least Squares Estimation / Maximum Likelihood Estimation – [1].
    - Poor Performance under Misspecification.
  - Kolmogorov-Smirnov Testing – [2].
    - Suffers massively from the Curse of Dimensionality, as do most IPM–adjacent techniques.
- Focused Techniques
  - 'Generalized Hosmer–Lemeshow Testing' – [3].
    - Procedurally complex.
  - Cameron–Trivedi Testing – [4].
    - Assesses for data dispersion that isn't unique to the Poisson Distribution.

Either heavily reliant on faulty heuristics or are overcomplex.

No generalisable direct comparison between measure allocation by the model and the measure underlying the observed data.

Newcastle University

Kernel Trickery, pt. 1

For more information – [5], [6], [7]

- Reproducing Kernel Hilbert Space $\mathcal{H}$ (RKHS)
  - Hilbert Space of functions $h : \mathcal{X} \to \mathbb{R}$ equipped with $\Phi_{\mathcal{H}} : \mathcal{X} \to \mathcal{H}$ such that:

$$\langle h, \Phi_{\mathcal{H}}(x) \rangle_{\mathcal{H}} = h(x) \; \forall \; x \in \mathcal{X}$$

- Reproducing Kernel of $\mathcal{H}$
  - $K : \mathcal{X}^2 \to [0, \infty)$ with:

$$K(x, y) = \langle \Phi_{\mathcal{H}}(x), \Phi_{\mathcal{H}}(y) \rangle_{\mathcal{H}} \; \forall \; x, y \in \mathcal{X}$$

    [8] proves that you can pick a K first and that as long as K is 'Symmetric and Positive Semi–Definite' then there is a RKHS with Reproducing Kernel $\mathcal{H}$.
- Kernel Mean Embedding (KME)
  - For $P \in \mathcal{P}(\mathcal{H}) = \left\{ Q \mid -\infty < \int h(x) dQ(x) < \infty \; \forall \; h \in \mathcal{H} \right\}$, μ_P is given by:

$$\mu_P = \int \Phi_{\mathcal{H}}(y) dP(y) \in \mathcal{H}$$

$$\langle \mu_P, \Phi_{\mathcal{H}}(x) \rangle_{\mathcal{H}} = \mu_P(x) = \int \langle \Phi_{\mathcal{H}}(x), \Phi_{\mathcal{H}}(y) \rangle_{\mathcal{H}} \, dP(y) = \int K(x, y) dP(y) \; \forall \; x \in \mathcal{X}$$

Kernel Trickery, pt. 2

For more information – [5], [6], [7]

- Double Kernel Mean Embedding (DKME)
  - For $P_1, P_2 \in \mathcal{P}(\mathcal{H}) = \left\{Q \mid -\infty < \int h(x) dQ(x) < \infty \; \forall \; h \in \mathcal{H}\right\}$:

$$\left\langle \mu_{P_1}, \mu_{P_2} \right\rangle_{\mathcal{H}} = \iint K(x,y) dP_1(x) dP_2(y) \; \forall \; x \in \mathcal{X}$$

- Characteristic Kernel
  - If the KME is Injective, then the underlying K is 'Characteristic'. The following two kernels are always Characteristic for Positive–Definite R and $n \in \mathbb{Z}_{\geq 0}$:

$$(x,y) \mapsto \exp\left(-\frac{1}{2}(x-y)^\top R^{-1}(x-y)\right), \; (x,y) \mapsto \frac{n!\left(\sum_{i=0}^{n} \frac{(n+i)!}{i!(n-i)!}\left(\sqrt{(8n+4)(x-y)^\top R^{-1}(x-y)}\right)^{n-i}\right)}{(2n)!\exp\left(\sqrt{(2n+1)(x-y)^\top R^{-1}(x-y)}\right)}$$

- Core Kernel of P over $\mathcal{H}$
  - For $P \in \mathcal{P}(\mathcal{H})$, the SPSD $K_P : \mathcal{X}^2 \to [0,\infty)$ with $K_P(x,y) = \langle \Phi_{\mathcal{H}}(x) - \mu_P, \Phi_{\mathcal{H}}(y) - \mu_P \rangle_{\mathcal{H}} \; \forall \; x, y$ satisfies $\int K_P(x,y) dP(y) = 0 \; \forall \; x$ and:

$$K_P(x,y) = K(x,y) - \int K(x,b) dP(b) - \int K(a,y) dP(a) + \iint K(a,b) dP(a) dP(b) \; \forall \; x, y$$

Newcastle University

Maximum Mean Discrepancy

- For some $\mathcal{F}$ with norm $\|\cdot\|_{\mathcal{F}}$ and $\mathcal{P}(\mathcal{F}) = \left\{Q \mid -\infty < \int f(x)dQ(x) < \infty \; \forall \, f \in \mathcal{F}\right\}$:

$$\text{MMD}_{\mathcal{F}}(P, Q) = \sup_{f \in \mathcal{F} \,:\, \|f\|_{\mathcal{F}} \leq 1} \left| \int f(x)dP(x) - \int f(x)dQ(x) \right| \; \forall \, P, Q \in \mathcal{P}(\mathcal{F})$$

- For a RKHS $\mathcal{H}$ with Characteristic Kernel $K : \mathcal{X}^2 \to (0, \infty)$, [7] show that:

$$\text{MMD}_{\mathcal{H}}(P, Q) = \|\mu_P - \mu_Q\|_{\mathcal{H}} = 0 \Leftrightarrow \mu_P = \mu_Q \Leftrightarrow P = Q$$

$$\text{MMD}_{\mathcal{H}}(P, R) = \|\mu_P - \mu_R\|_{\mathcal{H}} \leq \|\mu_P - \mu_Q\|_{\mathcal{H}} + \|\mu_Q - \mu_R\|_{\mathcal{H}} = \text{MMD}_{\mathcal{H}}(P, Q) + \text{MMD}_{\mathcal{H}}(Q, R)$$

- [9] show that:

$$\text{MMD}^2_{\mathcal{H}}(P, Q) = (\text{MMD}_{\mathcal{H}}(P, Q))^2 = \iint K_P(x, y)dQ(x)dQ(y)$$

Newcastle University

MMD between model class $P_\theta \in \mathcal{P}(\mathcal{H})$ and data $p_i \sim P^\dagger : 1 \leq i \leq N$ for $P^\dagger \in \mathcal{P}(\mathcal{H})$

$$0 \leq \mathrm{MMD}_{\mathcal{H}}\left(P_\theta, P^\dagger\right) \leq \mathrm{MMD}_{\mathcal{H}}\left(P_\theta, \frac{1}{N}\sum_{i=1}^{N}\delta_{p_i}\right) + \underbrace{\mathrm{MMD}_{\mathcal{H}}\left(\frac{1}{N}\sum_{i=1}^{N}\delta_{p_i}, P^\dagger\right)}_{\approx 0 \text{ as } N\to\infty \text{ from [10]}}$$

$$\mathrm{MMD}^2_{\mathcal{H}}\left(P_\theta, \frac{1}{N}\sum_{i=1}^{N}\delta_{p_i}\right) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}K_{P_\theta}\left(p_i, p_j\right)$$

We can find $P_\theta \approx P^\dagger$ by taking $N \to \infty$ and minimising $\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}K_{P_\theta}\left(p_i, p_j\right)$.

Newcastle University

MMD between $P_\theta \in \mathcal{P}(\mathcal{Y}|\mathcal{X})$ and $(x_i, y_i) \sim P^\dagger|P_\mathcal{X} : 1 \leq i \leq N$ for $P^\dagger \in \mathcal{P}(\mathcal{Y}|\mathcal{X})$ using [11].

$$0 \leq \text{MMD}_{\mathcal{H}_\mathcal{Z}}\left(P_\theta|P_\mathcal{X}, P^\dagger|P_\mathcal{X}\right) \leq \text{MMD}_{\mathcal{H}_\mathcal{Z}}\left(P_\theta|\frac{1}{N}\sum_{i=1}^{N}\delta_{x_i}, \frac{1}{N}\sum_{i=1}^{N}\delta_{(x_i,y_i)}\right) + \underbrace{\text{MMD}_{\mathcal{H}_\mathcal{Z}}\left(\frac{1}{N}\sum_{i=1}^{N}\delta_{(x_i,y_i)}, P^\dagger|P_\mathcal{X}\right)}_{\approx 0 \text{ as } N\to\infty \text{ from [10]}}$$

$$\left[\underbrace{\mathcal{X}\times\mathcal{Y}=\mathcal{Z}}_{\text{Joint Sample Space}} \mid \underbrace{K_\mathcal{X} \to \mathcal{H}_\mathcal{X}}_{\text{Covariate RKHS}} \mid \underbrace{K_\mathcal{X}\otimes K_\mathcal{Y} \to \mathcal{H}_\mathcal{X}\otimes\mathcal{H}_\mathcal{Y} = \mathcal{H}_\mathcal{Z}}_{\text{Joint RKHS}}\right] \qquad + \underbrace{C_\circ \text{MMD}_{\mathcal{H}_\mathcal{X}}\left(\frac{1}{N}\sum_{i=1}^{N}\delta_{x_i}, P_\mathcal{X}\right)}_{\approx 0 \text{ as } N\to\infty \text{ from [10]}}$$

$$\text{MMD}^2_{\mathcal{H}_\mathcal{Z}}\left(P_\theta|\frac{1}{N}\sum_{i=1}^{N}\delta_{x_i}, \frac{1}{N}\sum_{i=1}^{N}\delta_{(x_i,y_i)}\right) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} K_{P_\theta|\frac{1}{N}\sum_{k=1}^{N}\delta_{x_k}}\left((x_i,y_i),(x_j,y_j)\right)$$

Newcastle University

Next Steps

- Consistently calculating $\text{MMD}^2_{\mathcal{H}_\mathcal{Z}}\left(P_\theta | \frac{1}{N}\sum_{i=1}^{N}\delta_{x_i}, \frac{1}{N}\sum_{i=1}^{N}\delta_{(x_i,y_i)}\right)$
  - $\int K(x,y)dP(y)$, $\int\int K(x,y)dP(x)dQ(y)$ range from difficult to calculate perfectly to impossible to calculate perfectly.
  - Novel contributions to the canon in tracking some of these down.
- Developing a testing framework
  - Work with $T_N = N \min_{\theta\in\Theta} \text{MMD}^2_{\mathcal{H}_\mathcal{Z}}\left(P_\theta | \frac{1}{N}\sum_{i=1}^{N}\delta_{x_i}, \frac{1}{N}\sum_{i=1}^{N}\delta_{(x_i,y_i)}\right)$.
  - Leverage non-parametric neighbours of χ^2 testing and McDiarmid's Inequality.

Newcastle University

Poisson Gaussian KMEs

For $\lambda, \mu \geq 0$, $\rho > 0$, P_λ distributing $\mathbb{R}$ according to $\text{Po}(\lambda)$ and $K(x,y) = \exp\left(-\frac{1}{2}\left(\frac{x-y}{\rho}\right)^2\right)$, the following holds:

1. $\int K(x,y)dP_\lambda(y)$ is equal to:

$$\frac{\int_{-\infty}^{\infty} \frac{\exp\left(-\frac{1}{2}t^2\right)}{\sqrt{2\pi}} \left(\exp\left(\lambda \exp\left(\frac{x}{\rho^2}\right)\cos\left(\frac{t}{\rho}\right)\right)\cos\left(\lambda \exp\left(\frac{x}{\rho^2}\right)\sin\left(\frac{t}{\rho}\right)\right)\right) dt}{\exp\left(\lambda + \frac{x^2}{2\rho^2}\right)} \quad \forall\, x.$$

2. $\int\int K(x,y)dP_\lambda(x)dP_\mu(y)$ is equal to:

$$\frac{\int_{-\infty}^{\infty} \frac{\exp\left(-\frac{1}{2}t^2\right)}{\sqrt{2\pi}} \left(\exp\left((\lambda+\mu)\cos\left(\frac{t}{\rho}\right)\right)\cos\left((\lambda-\mu)\sin\left(\frac{t}{\rho}\right)\right)\right) dt}{\exp(\lambda+\mu)}.$$

Newcastle University

Rundown of the Proposed Framework

$H : \mathbb{R} \to \{0, 1\}$ is the Heaviside Step Function, and $k > 0$ bounds $K_{\mathcal{X}} \otimes K_{\mathcal{Y}}$ from above.

- Parametric Bootstrap Configuration.
  1. Resample (x_i, y_i) from trained model M times.
  2. For the m-th resample, swap out the old sample with the m-th resample in T_n and call it Δ_m.
  3. Calculate $\frac{1}{M} \sum_{m=1}^{M} H(\Delta_m - T_N)$
- Wild Bootstrap Configuration, based off [12].
  - Calculate the Gram Matrix of the Core at minimising parameters, then multiply it by N. Call this final matrix G_N.
  - Sample M-many N-dimensional random variables $\{W_m\}_{m=1}^{M}$ from either $U(\{-1, 1\}^N)$ or $\mathcal{N}(0, \mathbf{1}_N)$. Must all be from the same distribution though!
  - Calculate $\frac{1}{M} \sum_{m=1}^{M} H(W_m^\top G W_m - T_N)$
- Concentration Inequality Configuration, based off [13].
  - Calculate $\exp\left(-1 - \frac{\max\left\{\frac{T_N}{k}, 2\right\}}{2} + \sqrt{2 \max\left(\left\{\frac{T_N}{k}, 2\right\}\right)}\right)$

Different performances due to different amounts of collected info.

The END

Newcastle University

References I

[1] Richard M. Golden et al. “Consequences of Model Misspecification for Maximum Likelihood Estimation with Missing Data”. In: *Econometrics* 7 (2019). ISSN: 2225-1146. DOI: 10.3390/econometrics7030037. URL: https://www.mdpi.com/2225-1146/7/3/37.

[2] Donald W K Andrews. “A Conditional Kolmogorov Test”. In: *Econometrica* 65.5 (1997), p. 1097.

[3] Nikola Surjanovic, Richard Lockhart, and Thomas M. Loughin. *A Generalized Hosmer-Lemeshow Goodness-of-Fit Test for a Family of Generalized Linear Models*. 2021. arXiv: 2007.11049 [stat.ME]. URL: https://arxiv.org/abs/2007.11049.

[4] A.Colin Cameron and Pravin K. Trivedi. “Regression-based tests for overdispersion in the Poisson model”. In: *Journal of Econometrics* 46.3 (1990), pp. 347–364. ISSN: 0304-4076. DOI: https://doi.org/10.1016/0304-4076(90)90014-K. URL: https://www.sciencedirect.com/science/article/pii/030440769090014K.

Newcastle University

References II

[5] Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. en. 2008th ed. New York, NY: Springer, 2008.

[6] Bharath K. Sriperumbudur, Kenji Fukumizu, and Gert R. G. Lanckriet. "Universality, Characteristic Kernels and RKHS Embedding of Measures". In: *J. Mach. Learn. Res.* 12.null (July 2011), pp. 2389–2410. ISSN: 1532-4435.

[7] Arthur Gretton et al. "A Kernel Two-Sample Test". In: *J. Mach. Learn. Res.* 13 (2012), pp. 723–773. URL: https://api.semanticscholar.org/CorpusID:10742222.

[8] Nathan Aronszajn. "Theory of Reproducing Kernels.". In: *Transactions of the American Mathematical Society* 68 (1950), pp. 337–404. URL: https://api.semanticscholar.org/CorpusID:54040858.

[9] Chris. J. Oates. "Minimum Kernel Discrepancy Estimators". In: (2023). arXiv: 2210.16357 [stat.ME].

Newcastle University

References III

[10] Alain Berlinet and Christine Thomas-Agnan. *Reproducing Kernel Hilbert Space in Probability and Statistics*. Jan. 2004. ISBN: 978-1-4613-4792-7. DOI: 10.1007/978-1-4419-9096-9.

[11] P Alquier and M Gerber. "Universal robust regression via maximum mean discrepancy". In: *Biometrika* 111.1 (May 2023), pp. 71–92. ISSN: 1464-3510. DOI: 10.1093/biomet/asad031. eprint: https://academic.oup.com/biomet/article-pdf/111/1/71/56665593/asad031.pdf. URL: https://doi.org/10.1093/biomet/asad031.

[12] Oscar Key et al. "Composite Goodness-of-fit Tests with Kernels". In: (2024). arXiv: 2111.10275 [stat.ML].

[13] Francois-Xavier Briol et al. "Statistical inference for generative models with maximum mean discrepancy". In: (2019).