Approximate kernel regression in Stata using random Fourier features

Christopher James Rose

Norwegian Institute of Public Health

NIPH
CENTRE FOR EPIDEMIC INTERVENTIONS RESEARCH

Conflicts of interest

- I have no conflicts of interest to declare

NIPH
CENTRE FOR EPIDEMIC INTERVENTIONS RESEARCH

Outline

- Introduction: nonlinear regression and kernel regression
- Kernel methods and the random Fourier feature approximation
  - Not my own method: introduced by Rahimi and Recht (2007)
  - They won the 2017 NeurIPS Test of Time Award for the method
- I will introduce the `makerff` command, which has similar syntax to Stata's `makespline` command
- Example: approximate kernel Cox regression in Stata
- Conclusions

Introduction

What is nonlinear regression?

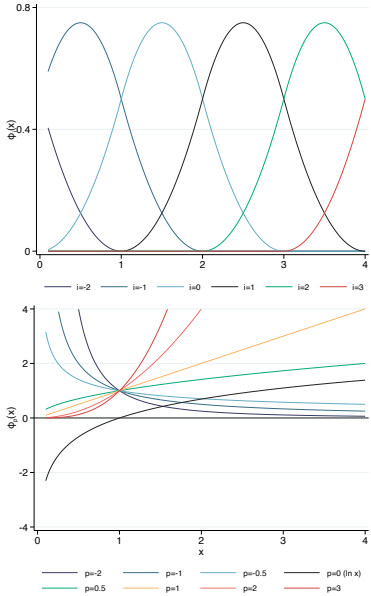
- Outcomes may depend on covariates via an unknown, nonlinear function $f(\mathbf{x})$
- Nonlinear methods adjust for confounders, model dose-response, predict risk, characterize treatment-effect heterogeneity
- Fractional polynomials, splines, kernel methods, random Fourier features take the form

$$f(x) = \sum_{j=1}^{k} \beta_j \, \phi_j(x)$$

coefficients

basis functions

- The methods differ in the form of ϕ_j and how they estimate β_j



Basis functions

- Splines (top), fractional polynomials (bottom)
- The analyst chooses the basis functions in advance
- A poorly chosen basis can underfit or misrepresent the true shape of *f*

NIPH
CENTRE FOR EPIDEMIC INTERVENTIONS RESEARCH

What does Stata already offer?

- `makespline` and `fp` generate a spline or polynomial basis for use with any estimator — but the user must choose the basis (knots, degree, powers)
- `npregress` fits fully nonparametric regression without a chosen basis — but for continuous outcomes only
- `h2oml` offers flexible ensemble/tree-based methods — but requires setting up and managing an external H2O cluster
- None of these model continuous, binary, count, and time-to-event outcomes without a user-specified basis at computational cost that scales linearly with sample size



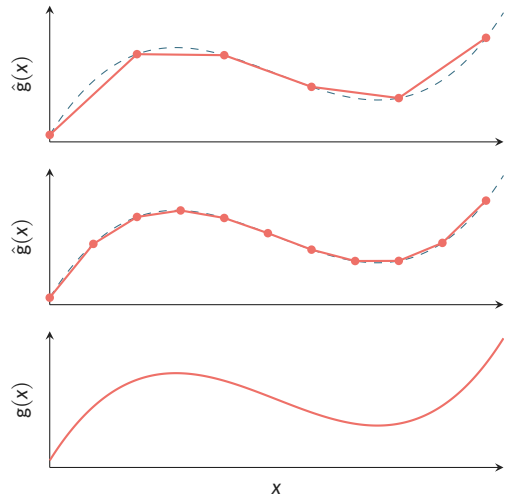
Kernel regression and the RFF approximation

Hilbert spaces: functions as vectors

- A function g can be approximated by a vector **g** of its values, indexed by points in its domain:

$$\mathbf{g} = \Big(\underset{0.5}{\mathrm{g}(0.5)},\ \underset{0.75}{\mathrm{g}(0.75)},\ \underset{1.0}{\mathrm{g}(1.0)},\ \ldots\Big)$$

- The approximation improves with the vector space's dimension
- In infinitely many dimensions, g is represented *exactly*: a **Hilbert space** $\mathbb{H}$
- An inner product in $\mathbb{H}$ gives functions lengths, angles, projections
- Kernel regression works directly in $\mathbb{H}$



Kernel regression and its fundamental limitation

- Kernel regression models f as linear in a Hilbert space $\mathbb{H}$ via feature map ϕ and an **inner product**:

$$\hat{f}(\mathbf{x}) = \left\langle \hat{\mathbf{w}},\ \phi(\mathbf{x}) \right\rangle_{\mathbb{H}}$$

coefficients in $\mathbb{H}$

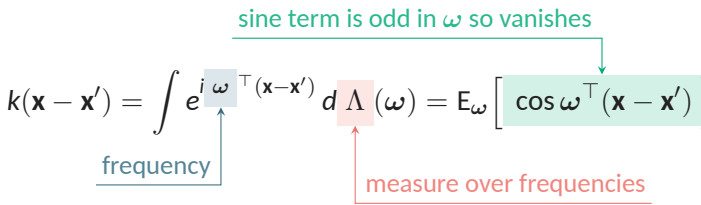
implicit, often infinite-dimensional feature map

- Not possible to compute directly in an infinite-dimensional space
- A kernel $k(\mathbf{x}, \mathbf{x}')$ lets us compute $\hat{f}$ for any $\mathbf{x}$ without ever forming $\phi(\mathbf{x})$
- Classical kernel regression evaluates k for every pair of observations, forming the $n \times n$ Gram matrix: storing it costs $O(n^2)$ memory, and decomposing it costs $O(n^3)$ time
- Impractical for large datasets

NIPH
CENTRE FOR EPIDEMIC INTERVENTIONS RESEARCH

Bochner's theorem and the Gaussian kernel

- A kernel is *shift-invariant* if it depends on $\mathbf{x}, \mathbf{x}'$ only through their difference, $k(\mathbf{x} - \mathbf{x}')$
- A positive definite k corresponds to a Hilbert-space inner product
- Bochner's theorem: k is positive definite iff it is the Fourier transform:



- When $k(\mathbf{0}) = 1$, Λ is a probability measure, so averaging over frequencies drawn from Λ approximates k
- If k is the Gaussian kernel, $\Lambda = N(\mathbf{0}, \sigma^{-2}\mathbf{I})$

NIPH
CENTRE FOR EPIDEMIC INTERVENTIONS RESEARCH

The random Fourier feature construction

- Cosine addition: $\cos \omega^\top (\mathbf{x} - \mathbf{x}') = \cos \omega^\top \mathbf{x} \cos \omega^\top \mathbf{x}' + \sin \omega^\top \mathbf{x} \sin \omega^\top \mathbf{x}'$
- Stack cos and sin, evaluated at q random frequencies, into a single vector $\mathbf{z}(\mathbf{x})$

the length-$2q$ random feature map

$$\mathbf{z}(\mathbf{x}) = \sqrt{\frac{1}{q}} \left(\cos \omega_1^\top \mathbf{x},\ \sin \omega_1^\top \mathbf{x},\ \ldots,\ \cos\ \omega_j\ ^\top \mathbf{x},\ \sin \omega_j^\top \mathbf{x} \right)$$

the jth frequency, drawn from Λ

number of random frequencies

- $\mathbf{z}$ is defined so that the product $\mathbf{z}(\mathbf{x})^\top \mathbf{z}(\mathbf{x}')$ is a mean over q random frequencies and approximates the expectation $\mathrm{E}_\omega[\cos \omega^\top (\mathbf{x} - \mathbf{x}')]$, and hence $k(\mathbf{x} - \mathbf{x}')$
- Approximation error shrinks at rate $O(1/\sqrt{2q})$ as q grows

NIPH
CENTRE FOR EPIDEMIC INTERVENTIONS RESEARCH

Ridge-penalized estimation to approximate kernel regression

- Without a penalty, $\mathbb{H}$'s many dimensions let $\hat{f}$ interpolate the data exactly (overfitting)
- Kernel methods control this using the Hilbert-space penalty $\|\hat{\mathbf{w}}\|^2_{\mathbb{H}}$
- In the approximation space, this corresponds to the ℓ_2 ridge penalty
- This leads to the familiar ridge regression estimator

negative log-likelihood function

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_i L\left(y_i, \mathbf{z}(\mathbf{x}_i)^\top \boldsymbol{\beta}\right) + \lambda \|\boldsymbol{\beta}\|^2$$

controls shrinkage

- In Stata: `elasticnet <model> y rff*, alpha(0)`
- As $q \to \infty$, this converges to exact kernel regression

NIPH
CENTRE FOR EPIDEMIC INTERVENTIONS RESEARCH

Example

Maternal education and inter-birth interval

- `cattaneo2` (Cattaneo 2010): 4,642 births, with 16 covariates including parental race, age, education, and prior infant deaths (convenient: ships with Stata)
- Q: is there a difference in inter-birth interval (`monthslb`) between mothers with lower versus higher educational attainment (`medu`)?
  - Operationalize the question by comparing the 10th and 90th percentiles of `medu` in terms of restricted mean survival time (RMST)
  - Estimate using approximate kernel Cox regression, bootstrapped g-computation
- This example is illustrative only, not intended to yield "impressive" or substantively correct results

Some sequences shortened; see the forthcoming paper for full code.

Constructing the random Fourier features

- Store the covariates in a local macro, and then run `makerff`, expanding the macro:

```
local spec bweight i.mmarried i.mrace i.frace i.mhisp i.fhisp i.foreign
           i.alcohol deadkids mage fage medu fedu i.prenatal
           i.lbweight i.fbaby

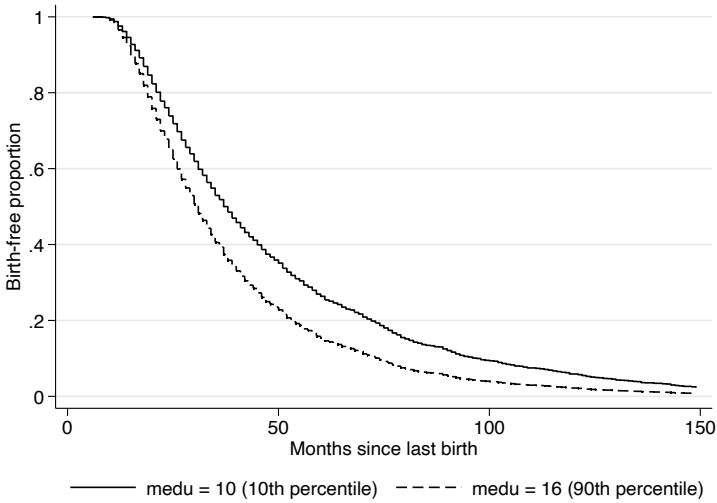
makerff `spec'
```

- Some of the 16 variables are factors; `makerff` maps the 28-dimensional design matrix to a $2q = 280$-dimensional space
- By default, `makerff` sets $q = \min(5p, \lfloor n/2 \rfloor)$: here $p = 28$ and $n = 4{,}642$, so $q = \min(140, 2321) = 140$

Approximate kernel Cox regression and g-computation

- Outcome: `monthslb` (inter-birth interval) among 2,397 multiparous mothers
- `elasticnet cox monthslb rff*, alpha(0)` (requires Stata $\geq$18)
- Coefficients on generated features are not directly interpretable
- Use g-computation (Hernán 2020) to answer Q:
  - Fit once and save the frequency matrix
  - Set `medu` to a counterfactual value (e.g., `medu` = 10, its 10th percentile) and regenerate the features
  - Predict individual hazard ratios
  - Estimate the marginal survivor curve under the counterfactual
- Do this for each counterfactual (10th and 90th percentiles of `medu`)

Marginal survivor curves



NIPH
CENTRE FOR EPIDEMIC INTERVENTIONS RESEARCH

Bootstrap inference

- Restricted mean survival time (RMST) can be used to estimate and compare expected inter-birth intervals (Royston 2011)
- RMST is the area under a survivor curve up to a restriction time (e.g., the last observed follow-up)
- The RMST difference has no closed-form standard error
- Estimate uncertainty with the nonparametric bootstrap (resample mothers, refit the kernel Cox model, repeat the g-computation)

The saved frequency matrix is reused, not redrawn, within each bootstrap replicate, so resampling mothers — not frequencies — drives the variation.

Results

RMST (months)		Difference	SE	*p*	95% CI
90th %-ile	10th %-ile	(months)			
32.92	43.20	-10.28	1.58	< 0.001	-13.4 to -7.2

- Expected inter-birth interval is about 33 months for better-educated mothers
- Expected inter-birth interval is about 43 months for less-educated mothers
- Better-educated mothers have a subsequent birth about 10 months sooner
- Unlikely to be explained by chance alone

NIPH
CENTRE FOR EPIDEMIC INTERVENTIONS RESEARCH

Conclusions

Conclusions

- `makerff` approximates kernel regression; no need to choose a basis
- Covers continuous, binary, count, and time-to-event outcomes
- Practical: scales linearly with sample size
- Only one real tuning parameter, q (approximation provably improves with q)
- Feature coefficients aren't interpretable; summarize using predictions or marginal effects
- Natural extension: additional kernel functions (rational quadratic, Matérn, …)
- Accepted for publication in *The Stata Journal*, to appear in 2027

References

- Cattaneo, Matias D. (2010). “Efficient semiparametric estimation of multi-valued treatment effects under ignorability”. In: *Journal of Econometrics* 155.2, pp. 138–154.
- Hernán, Miguel A. *et al.* (2020). *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC.
- Li, Zhu *et al.* (2021). “Towards a unified analysis of random Fourier features”. In: *Journal of Machine Learning Research* 22.108, pp. 1–51.
- Rahimi, Ali *et al.* (2007). “Random Features for Large-Scale Kernel Machines”. In: *NeurIPS*. Vol. 20. Curran Associates, Inc.
- — (2008). “Weighted Sums of Random Kitchen Sinks: Replacing Minimization with Randomization in Learning”. In: *NeurIPS*. Vol. 21. Curran Associates, Inc.
- Royston, Patrick *et al.* (2011). “The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials”. In: *Statistics in Medicine* 30.19, pp. 2409–2421.



Questions?

NIPH
CENTRE FOR EPIDEMIC
INTERVENTIONS RESEARCH

Additional slides

Common shift-invariant kernels

Kernel	$k(r)$	Best suited for
Gaussian	$\exp\left(-\frac{r^2}{2\ell^2}\right)$	Very smooth functions
Rational quadratic	$\left(1+\frac{r^2}{2\alpha\ell^2}\right)^{-\alpha}$	Functions with multiple length scales
Exponential	$\exp\left(-\frac{r}{\ell}\right)$	Rough, non-smooth functions
Matérn	$\frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\sqrt{2\nu}\,r}{\ell}\right)^{\nu} K_\nu\left(\frac{\sqrt{2\nu}\,r}{\ell}\right)$	Functions with tunable smoothness
Periodic	$\exp\left(-\frac{2\sin^2(\pi r/p)}{\ell^2}\right)$	Periodic or seasonal patterns

$r = \|\mathbf{x}-\mathbf{x}'\|$; ℓ is a lengthscale, α a scale-mixture parameter, ν a smoothness parameter, p a period, and K_ν the modified Bessel function of the second kind.

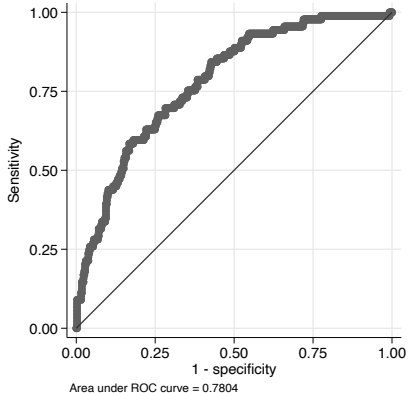
NIPH
CENTRE FOR EPIDEMIC INTERVENTIONS RESEARCH

Additional example: kernel logistic regression for risk prediction

Predicting maternal smoking during pregnancy, using the `cattaneo2` data

- Set aside a 10% holdout up front, for out-of-sample evaluation
- Uses the same random Fourier features (`rff*`) built earlier for the kernel Cox example
- `elasticnet logit y rff*, alpha(0)` fit on the estimation sample, ridge penalty chosen by 10-fold cross-validation
- Post-estimation: predict on the holdout set and assess predictive performance via ROC analysis

Additional example: results



AUC $= 0.78$: about a 78% chance a smoker is predicted to have a higher probability of smoking than a non-smoker in a randomly chosen pair

NIPH
CENTRE FOR EPIDEMIC INTERVENTIONS RESEARCH

The parallelogram law

$$\|\mathbf{u}\|_\gamma = \left(\sum_i |u_i|^\gamma \right)^{1/\gamma}$$

$\gamma = 1$: lasso; $\gamma = 2$: ridge

- A norm is induced by an inner product only if it obeys the parallelogram law: $\|\mathbf{u} + \mathbf{v}\|_\gamma^2 + \|\mathbf{u} - \mathbf{v}\|_\gamma^2 = 2\|\mathbf{u}\|_\gamma^2 + 2\|\mathbf{v}\|_\gamma^2$ for all $\mathbf{u}, \mathbf{v}$
- Among the ℓ_γ norms, only ℓ_2 satisfies this
- A lasso-type (ℓ_1) penalty would remove individual random frequencies from $\mathbf{z}(\mathbf{x})$, rather than shrinking the fitted function within a fixed basis
- This is why `elasticnet` needs `alpha(0)` — any ℓ_1 component breaks the correspondence with kernel regression

NIPH
CENTRE FOR EPIDEMIC INTERVENTIONS RESEARCH

Convergence guarantees by outcome type

Model	$L(y, \eta)$	Bounded deriv.	Guarantee
Linear	$(y-\eta)^2/2$	No	Formal
Logistic	$\log(1+e^{\eta}) - y\eta$	Yes	Formal
Probit	$-\log \Phi(s\eta)$	No	Informal (tail growth)
Poisson	$e^{\eta} - y\eta$	No	Informal (tail growth)
Cox	Partial likelihood	—	Informal (structural)

$s = 2y - 1$; Φ is the standard normal cdf.

- As $q \to \infty$, the ridge-penalized RFF estimator converges to the exact kernel-regularized estimator, for any log-likelihood with bounded derivative (Rahimi 2008; Li 2021)
- A bounded derivative limits how approximation error in η affects the loss; bounded features ($\|\mathbf{z}(\mathbf{x})\| = 1$) and the ridge penalty confine η similarly otherwise
- The logistic log-likelihood has bounded derivative in η — the correspondence with exact kernel regression is formal; so is the linear case
- Probit, Poisson, and Cox are not covered — but bounded features and the ridge penalty keep η well-behaved in practice

NIPH
CENTRE FOR EPIDEMIC INTERVENTIONS RESEARCH