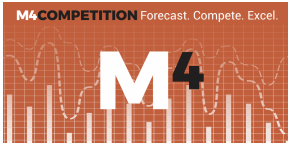
Nonparametric Value-at-Risk

Some guidance for the use of Neural Nets

Philipp Ratz

CFE London 2022

From theory - to practice



Source: [Institute for the future, 2018]

Research and practice seem to diverge on ML: summarised in the results of the *M4 Forecasting Competition* [Makridakis et al., 2020]

"the most accurate ML method was less accurate than the worst statistical ones"

"the top three [...] introduced information from multiple series "

Goals

1. If there is a discrepancy between theory and practice can we explain why?

Goals

1. If there is a discrepancy between theory and practice can we explain why?
   - Need some statistical tools that go beyond the capability

Goals

1. If there is a discrepancy between theory and practice can we explain why?

- Need some statistical tools that go beyond the capability

2. Can we identify situations where using more complex (harder to fit, harder to control, harder to interpret..) models makes sense?

A general framework

Enter sieve estimation - this allows us to consider neural networks as nonparametric estimators. Consider the general framework (based on excellent chapter from [Chen, 2007])

$$y = f(x) + \varepsilon$$

where f is some function of interest, to be estimated

Parametric approach

$$\mathcal{P} = \{f_\theta | \theta \in \Theta\}$$

and $\Theta \subseteq \mathbb{R}^d$

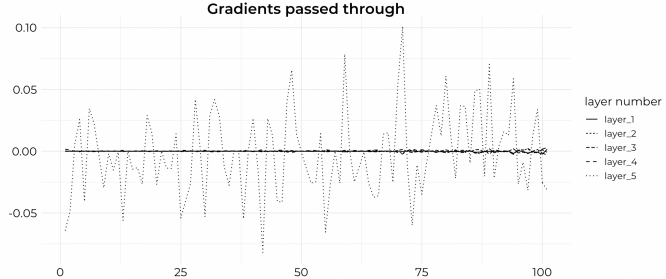
Semi-(Non)parametric

$$\mathcal{P} = \{f_\theta | \theta \in \Theta\}$$

and Θ is infinite dimensional

The gradient issue

Most of the asymptotic theory for single-hidden layer ANN was derived in the late 90s/early 2000s - they used squashing functions. But for the sigmoid $\frac{d\sigma(x)}{dx} = \sigma(x)(1-\sigma(x)) \leq 0.5$



ReLU$(x) = \max\{0, x\}$ suffers less from this.

Goal of the investigation

1. Revisit and expand previous work
   - Extend rates to include ReLU
   - Revisit the "hybrid" approach
2. Study finite sample properties (and compare to other nonparametric approaches)
3. Application - conditional VaR

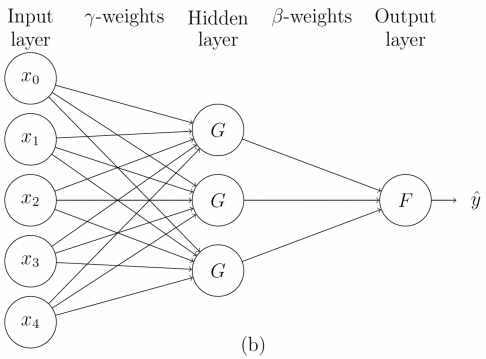
A theoretical framework I

The famed *universal approximation theorem* by for example [Hornik et al., 1989] tells us that it is possible to approximate almost any f with a neural network.

$$\hat{y}_i = \sum_{k=1}^{r_n} \beta_k g(\gamma_k^\top x_i)$$

But this is valid only in the limit (of width or depth). How can we use this to derive more statistical properties?

A theoretical framework II



Throughout, we assume the vector of covariates is in $\mathbb{R}^d$, $y \in \mathbb{R}$ and we have a random sample of size n

A theoretical framework III

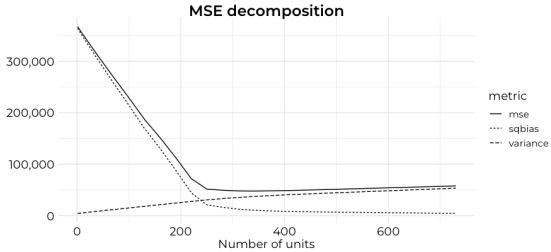
Θ is an infinite dimensional parameter space with metric ρ and there exists a population criterion function $Q : \Theta \to \mathbb{R}$ and its empirical counterpart $\hat{Q}_n : \Theta \to \mathbb{R}$, usually in the form of a loss

$$\hat{\theta} = \arg\max_{\theta} -\frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(\theta, x_i, y_i)$$

Issue: Θ is infinite dimensional! So $Q(\theta_0) - Q(\theta_k) \to 0$ but $\rho(\theta_0, \theta_k) \not\to 0$
Instead we can define a series of approximation spaces $\Theta_n \subseteq \Theta_{n+1} \subseteq \cdots \subseteq \Theta$ and maximize $\hat{Q}_n$ over this simpler approximation spaces.

Sieve approximation with hidden units

Recall: $\text{MSE} = \mathbb{E}[(y - \hat{f})^2] = \text{Bias}^2(\hat{f}) + \text{Var}[\hat{f}] + \sigma^2$



$$y_i = 0.4(x_i - 10)^3 + \sin(2x_i) + \varepsilon_i, \quad x_i \sim \mathcal{U}[-10, 10]$$

Sidenote: There is some discussion about that [Neal et al., 2018]

Theoretical Results I

Assume (among other conditions)

1. The function we are trying to approximate is continuous (this may be relaxed slightly)
2. The sieve space is compact
3. The absolute value of the weights within the network is bounded above
4. Within a neighbourhood around θ_0, the loss function is smooth for $\theta \in \Theta$

Theoretical Results II

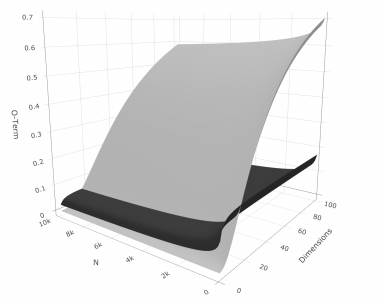
Then similar to the approach of [Chen, 2007] we can derive the rate of convergence for the neural network sieve

Proposition

The rate of convergence for a sieve estimatior with ReLU activation functions is $\mathcal{O}(\frac{n}{\log[n]}^{-(1+\frac{2}{d+1})/4(1+\frac{1}{1+d})})$. This is the same rate obtained by others such as [Chen and Shen, 1998, Chen, 2007]

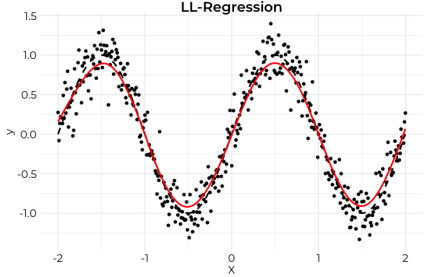
Theoretical Results III

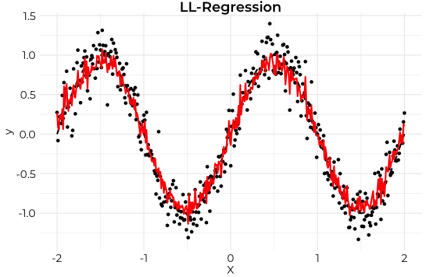
Why does this matter? Compare to convergence rate of density based kernel estimator



Adding noise to the equation I

A simple simulation example:





$$y = \sin[\pi x_1] + \varepsilon$$
$$\varepsilon \sim \mathcal{N}(0, 0.2^2), x_1 \in [-2, 2]$$

$$\hat{y} = LL(x_1, \ldots, x_6)$$
$$x_2 - x_6 \sim \mathcal{N}(0, 2^2)$$

Adding noise to the equation II

Instead, consider a slightly more complex simulation setup:

$$y_i = \sum_{l=1}^{k} f_l(x_{i,l}) + \varepsilon_i, \quad i = 1, \ldots, 500$$

with f_l different functions of an input variable and $\nu \sim \mathcal{N}(0, 7^2)$

Adding noise to the equation III

		Relevant Predictors		
	Noise	2	5	15
Nonparametric (LL)	1	6.11*	18.31*	59.13
	10	28.98	46.48	79.30
	20	44.02	66.65	97.99
ANN	1	22.33	33.99	41.86*
	10	27.10*	36.85*	47.03*
	20	34.87*	43.27*	51.72*

Table: Mean Squared prediction Error of ANN and KD estimators for varying complexity of functions. The star (*) denotes the lowest MSPE value across the models

An application for Risk management? I

The objective is flexible, hence it also holds for quantile regression!

Let Y be a real valued random variable with $F_Y(y) = \mathbb{P}[Y \leq y]$, define the α quantile as: $q_Y(\alpha) = F_Y^{-1}(\alpha)$. Further, for some covariates x define the conditional quantile as $\mathbb{E}[\mathbb{1}_{Y \leq g(x)}|x] = \alpha$ and we can then approximate the quantile $\hat{y} = g(x)$ with a neural network sieve. Then we can specify the quantile loss as:

$$\rho_\alpha(y - \hat{y}) = (\alpha - 1) \sum_{y_i < \hat{y}} (y_i - \hat{y}_i) + \alpha \sum_{y_i \geq \hat{y}} (y_i - \hat{y})$$

An application for Risk management? II

A natural starting point for risk management: A model where the quantile is measured directly: Consider the CAViaR Model as introduced by [Engle and Manganelli, 2004].

$$\text{VaR}_t = \beta_1 \text{VaR}_{t-1} + \beta_2 y_{t-1}$$

where y_{t-1} can be a vector, of information available at time $t-1$.

A semiparametric specification I

Proposition

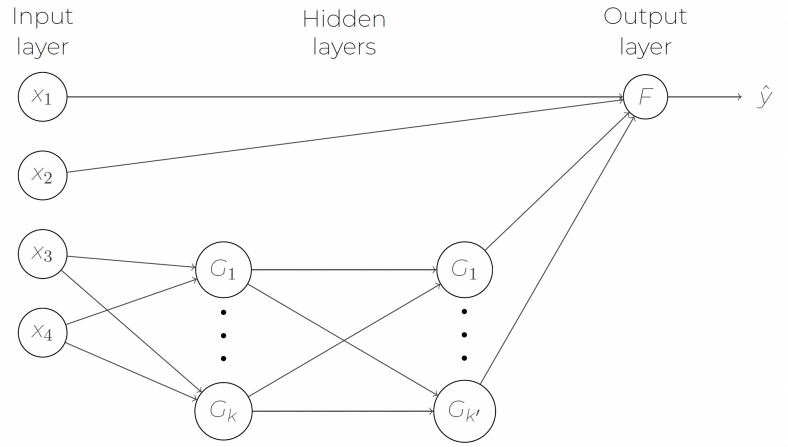
By following [Greene, 2018] we can estimate the parametric part from a semi-parametric specification of a neural network sieve with $\hat{\beta} = (X' M_{\mathcal{G}} X) X' M_{\mathcal{G}} Y$

Instead of the standard caviar model, consider a semiparametric version thereof:

$$f_t(\theta) = \sum_{j=1}^{p} \beta_j f_{t-j}(\theta) + \phi(x_{t-1})$$

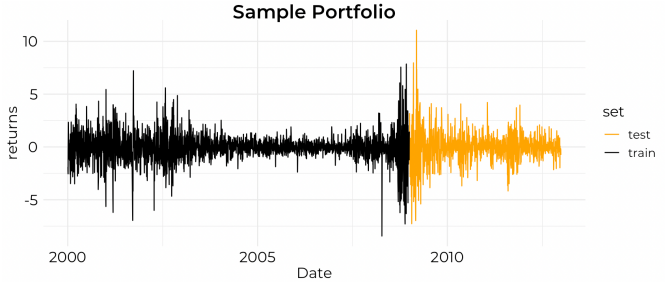
with ϕ some possibly nonlinear function of past returns and other covariates.

A semiparametric specification II



A portfolio approach - with some data I

Take data from Stockmarket, Forex and Commodities from 2000 - 2014. Generate random weights (to simulate different portfolios) and evaluate.



A portfolio approach - with some data II

	GARCH(2,1)	ANN	MGARCH(1,1)	MANN
Exceedances	14	11	23.02	10.56
Failures Test	0.23	0.75	0.52	0.02
Duration Test	0.15	0.01	0.12	0.06
Change $\int \text{VaR}_\alpha$	-27.7%	-13.8%	-25.0%	-6.1%
Coef. VaR_{t-1}		0.66		0.46
Std. Error		0.19		0.11

Table: *Predictive performance across univariate and multivariate models. For the tests of the multivariate models the fraction of rejected portfolios is reported (eg. 0.02 - 2% of portfolios were rejected by the test)*

Multivariate approach I

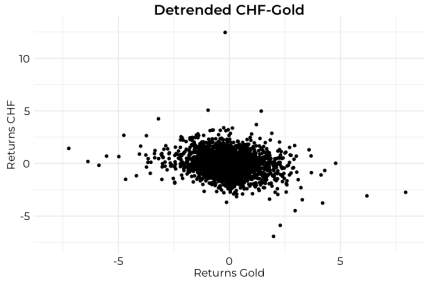
Recall: the definition of the sieve estimator is extremely flexible!

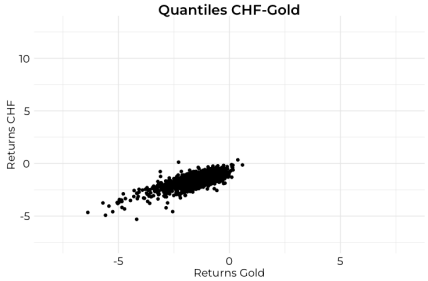
Proposition

The multivariate prediction problem $y = f(x) + \varepsilon$, where $y \in \mathbb{R}^K$ can be approximated with a neural network sieve

This also extends to quantile regression problems, and remember the conclusions from [Makridakis et al., 2020]

Multivariate approach II





Take-away points

- "Old" results are still valid for "new" activations
- Consider using neural networks as nonparametric estimators if you have:
  - Many covariates
  - Perhaps some redundant columns
  - Reasonable n
- For multivariate problems there might be a simple solution that is more robust (current research)

Appendix

Image on title page: Montréal from Kondiaronk Belvedere

References I

Chen, X. (2007).
Large sample sieve estimation of semi-nonparametric models.
Handbook of econometrics, 6:5549–5632.

Chen, X. and Shen, X. (1998).
Sieve extremum estimates for weakly dependent data.
Econometrica, pages 289–314.

Engle, R. F. and Manganelli, S. (2004).
Caviar: Conditional autoregressive value at risk by regression quantiles.
Journal of business & economic statistics, 22(4):367–381.

Greene, W. H. (2018).
Econometric analysis.

Hornik, K., Stinchcombe, M., and White, H. (1989).
Multilayer feedforward networks are universal approximators.
Neural networks, 2(5):359–366.

References II

- Institute for the future (2018).
  M4 forecasting competition.
- Makridakis, S., Spiliotis, E., and Assimakopoulos, V. (2020).
  The m4 competition: 100,000 time series and 61 forecasting methods.
  International Journal of Forecasting, 36(1):54–74.
- Neal, B., Mittal, S., Baratin, A., Tantia, V., Scicluna, M., Lacoste-Julien, S., and Mitliagkas, I. (2018).
  A modern take on the bias-variance tradeoff in neural networks.
  arXiv preprint arXiv:1810.08591.