



AgEcon SEARCH
RESEARCH IN AGRICULTURAL & APPLIED ECONOMICS

The World's Largest Open Access Agricultural & Applied Economics Digital Library

This document is discoverable and free to researchers across the globe due to the work of AgEcon Search.

Help ensure our sustainability.

Give to AgEcon Search

AgEcon Search
<http://ageconsearch.umn.edu>
aesearch@umn.edu

*Papers downloaded from **AgEcon Search** may be used for non-commercial purposes and personal study only. No other use, including posting to another Internet site, is permitted without permission from the copyright owner (not AgEcon Search), or as allowed under the provisions of Fair Use, U.S. Copyright Act, Title 17 U.S.C.*

Semiparametric panel data models using neural networks

Andrew Crane-Droesch*

May 11, 2017

Abstract

This paper presents an estimator for semiparametric models that uses a feed-forward neural network to fit the nonparametric component. Unlike many methodologies from the machine learning literature, this approach is suitable for longitudinal/panel data. It provides unbiased estimation of the parametric component of the model, with associated confidence intervals that have near-nominal coverage rates. Simulations demonstrate (1) efficiency, (2) that parametric estimates are unbiased, and (3) coverage properties of estimated intervals. An application section demonstrates the method by predicting county-level corn yield using daily weather data from the period 1981-2015, along with parametric time trends representing technological change. The method is shown to out-perform linear methods such as OLS and ridge/lasso, as well as random forest. The procedures described in this paper are implemented in the R package `pane1NNET`.¹

Neural networks[1] (termed “deep learning”[2] in some contexts) are the current state-of-the-art in machine learning and artificial intelligence. They have been successfully applied to tasks ranging from computer vision, natural language processing, self-driving cars, and quantitative finance. Hardware and software constraints limited the usefulness of these computationally-intensive approaches for many years. However, recent advances in hardware and software – and especially in algorithm design – have made them increasingly accessible and effective.

Ultimately however, neural networks are nothing more than algorithms for finding an optimal set of derived regressors from a set of input variables. These derived regressors are then used in a linear model of some outcome. Given that neural networks ultimately yield linear models, many standard econometric techniques can be applied to them in a fairly straightforward manner. That insight forms the crux of this paper.

In particular, this paper presents a semi-parametric extension to cross-sectional and panel data models, using neural networks to fit the nonparametric component of the model. The model discussed here is quite general, and is variously applicable to high-dimensional regression adjustment,

*Economic Research Service, US Department of Agriculture. andrew.crane-droesch@ers.usda.gov. The views expressed are those of the author and should not be attributed to the Economic Research Service or USDA

¹<https://github.com/cranedroesch/pane1NNET>

heterogeneous/individualized treatment effect estimation, general nonparametric regression problems, and forecasting with longitudinal data.

The basic model is

$$y_{it} = \alpha_i + \mathbf{X}_{it}\beta + f(\mathbf{Z}_{it}) + \epsilon_{it} \quad (1)$$

where y is an outcome for unit i in time t^2 , \mathbf{X} is a $N \times P_X$ matrix of data to be represented linearly, and \mathbf{Z} is a $N \times P_Z$ matrix of variables to be represented nonparametrically. Choices of which variables to be included in \mathbf{X} or \mathbf{Z} are left to the modeller. As will be made clear below, it will be appropriate to include variables in \mathbf{X} where the modeler has knowledge of appropriate parametric structure and/or desires unbiased marginal effects. All or some elements of \mathbf{Z} may be included in \mathbf{X} , yielding a linear “main effect” along with a nonlinear component. The compound error $\alpha_i + \epsilon_{it}$ represents between-unit and within-unit variability, respectively.

Rather than estimating hundreds or thousands of individual effect parameters, standard econometric practice removes α via the “within” transformation:

$$y_{it} - \bar{y}_i = (\mathbf{X}_{it} - \bar{\mathbf{X}}_i)\beta + f^*(\mathbf{Z}_{it} - \bar{\mathbf{Z}}_i) + \epsilon_{it} - \bar{\epsilon}_i \quad (2)$$

This fails however when $f()$ is nonlinear, because $f(\mathbf{Z}_{it}) \neq f^*(\mathbf{Z}_{it} - \bar{\mathbf{Z}}_i)$. While multidimensional linear basis expansions of $f(\mathbf{Z})$ can solve this problem, they are quickly overcome by the curse of dimensionality as P_Z grows.

Other machine learning approaches are also ill-suited to estimating 1. Random forests[3] can consistently[4, 5] and efficiently estimate models of the form $y = f(\mathbf{Z}) + \epsilon$, but they have no clear way to incorporate known parametric structure where available, nor an obvious way to eliminate the nuisance parameters α . Elastic-net regression[6] and generalized additive models[7] can handle fixed effects by first projecting-out fixed effects, but they require extremely large multidimensional basis expansions if they are to serve as universal approximators of an arbitrary $f()$ with a high-dimensional \mathbf{Z} .

This paper contributes to the econometric literature on nonparametric fixed effects models, which have heretofore relied on series estimation or kernel methods[8, 9, 10] and thereby suffer greatly from the curse of dimensionality. This paper also contributes to the machine learning literature, which has not yet developed a nonparametric algorithm suited to settings with repeated observations of many cross-sectional units. Finally, this paper contributes to the emerging literature at the intersection of machine learning and causal inference[11], particularly that which seeks to estimate personalized treatment effects[5, 12].

²The individual-time indices are presented without loss of generality. Any context with repeated observations of one or more cross-sectional unit is admissible in this framework.

1 Neural networks

1.1 Panel data

This paper proposes an estimator of 1 using a feed-forward neural network for the nonparametric part:

$$\begin{aligned} y_{it} &= \alpha_i + \mathbf{X}_{it}\beta + \mathbf{V}_{it}^1\Gamma^1 + \epsilon_{it} \\ \mathbf{V}_{it}^1 &= a(\gamma^2 + \mathbf{V}_{it}^2\Gamma^2) \\ \mathbf{V}_{it}^2 &= a(\gamma^3 + \mathbf{V}_{it}^3\Gamma^3) \\ &\vdots \\ \mathbf{V}_{it}^L &= a(\gamma^L + \mathbf{Z}_{it}\Gamma^L) \end{aligned} \tag{3}$$

where \mathbf{V}^l are derived variables or “nodes” at the l th layer. The parameters Γ^L (termed “weights” in the computer science literature) map the data \mathbf{Z} to the outcome via the intermediate nodes. *A priori*, no particular interpretation is attached to the derived variables \mathbf{V} – they are simply nonlinear combinations of \mathbf{Z} that are chosen to make the model fit well.

The number of layers and the number of nodes per layer (i.e.: the dimension of \mathbf{V}^l) is a hyperparameter chosen by the modeler. Note that $\Gamma^{2:L}$ are matrices of dimension equal to the number of nodes of the l th layer and the next layer up. The function $a(\cdot)$ is termed the “activation” function, and maps V from the real line to a defined interval – common choices are sigmoids such as the logistic and the hyperbolic tangent. More recently, the “rectified linear unit” [13] (and variants [14, 15]) activation function – $a(x) = \max(0, x)$ – has been shown to improve performance, especially in networks with many layers. Hornik et al. [16] have shown that neural networks can approximate any continuous function, given sufficiently many nodes and/or layers.

For a neural network with two layers, 3 can be written more compactly as

$$y_{it} = \alpha_i + \mathbf{X}_{it}\beta + a(a(\mathbf{Z}_{it}\Gamma^3)\Gamma^2)\Gamma^1 + \epsilon_{it} \tag{4}$$

Because the top layer is a linear model in the \mathbf{X} ’s and the derived variables \mathbf{V}^1 , the “within” transformation³ $y_{it} - \bar{y}_i = (\mathbf{X}_{it} - \bar{\mathbf{X}}_i)\beta + (\mathbf{V}_{it}^1 - \bar{\mathbf{V}}_i^1)\Gamma^1 + \epsilon_{it} - \bar{\epsilon}_i$ holds without altering β or Γ^1 . Fixed effects are thus removed.

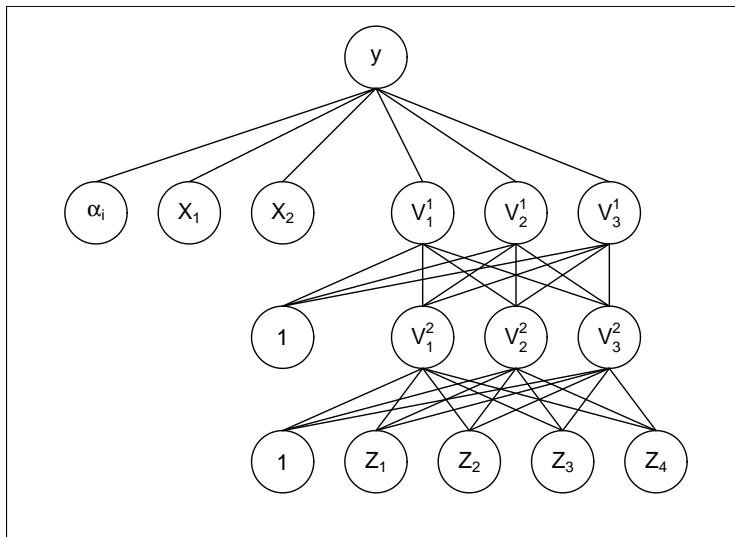
1.2 Estimation

The above are special cases of neural networks for continuous variables, and are fit to data in the same way as the general case. The loss surface for neural networks is generally nonconvex, and methods for minimizing this loss have been the focus of substantial research⁴. Most methods

³This is without loss of generality; multi-way fixed effects can similarly be projected-out without altering the top-level parameters.

⁴See Friedman et al. [17] for an introduction, and Ruder [18] for a literature review

Figure 1: Schematic drawing corresponding to equation 3, for a model with two parametric covariates, four nonparametric covariates, and two hidden layers with three nodes each. Line segments represent parameters, circles represent variables or derived variables.



are variants of minibatch gradient descent, which iteratively calculates the derivatives of the loss function with respect to the parameters for a subset of the data, alters the parameter estimates in those directions according to a step size, and proceeds iteratively with a new sample of data to convergence. This approach is popular in large-scale predictive applications because it works with very large datasets, including those that can not fit into a computer’s memory. Quasi-Newton methods are also feasible and work well with smaller and less complex networks.

In general, neural networks are overparameterized and will overfit the data, leading to poor performance when generalizing to new data. One way of controlling this is to tune the number of nodes and layers. Alternatively, the number of nodes and layers can be chosen to be sufficiently high as to ensure a good fit, and the parameters then chosen to minimize the penalized loss function $R = N^{-1} \sum_i (y_i - \hat{y}_i) + \lambda \sum \theta^2$ where θ is a vector of each of the parameters in the model, from each layer. Because regularization induces bias, parameters for which unbiased estimates are desired should not be included in θ – this may apply to the coefficients associated with parametric terms \mathbf{X} , if unbiased estimates of marginal effects are desired. The tuning parameter λ can then be chosen by some variant of cross-validation, including optimization against a withheld test set where k -fold cross-validation is impractical due to computational constraints.

1.2.1 The “OLS trick”

While gradient descent methods generally perform well, they are inexact. The top level of a neural network is a linear model however, in derived regressors in the typical context or in a mixture of derived regressors and parametric terms in our semiparametric context. Ordinary least squares provides a closed-form solution for the parameter vector that minimizes the unpenalized loss function, given a set of derived regressors, while ridge regression provides an equivalent solution for the penalized case.

In particular, after the m th iteration, the top level of the network is

$$y_{it} = \alpha_i + \mathbf{X}_{it}\beta^m + \mathbf{V}_{it}^m\Gamma^m + \epsilon_{it}$$

Because gradient descent is inexact, the parameter sub-vector $[\beta^m, \Gamma^m] = \theta^m$ does not satisfy

$$\min_{\theta} (\mathbf{y}^{dm} - \mathbf{W}^{dm}\theta)^T (\mathbf{y}^{dm} - \mathbf{W}^{dm}\theta) + \tilde{\lambda}\theta^T\theta \quad (5)$$

where $\mathbf{W} = [X, V]$, dm indicates the “within” transformation and $\tilde{\lambda} > \lambda$ is the penalty corresponding to the “budget” that is “left over” after fitting the lower level parameters, which generate V . Given that the general ridge regression solution is equivalent to minimizing $((\mathbf{y} - \mathbf{W}\theta)^T (\mathbf{y} - \mathbf{W}\theta))$ such that $\theta^T\theta < c$, one may calculate the implicit $\tilde{\lambda}$ for the top level of the neural network by minimizing

$$\min_{\lambda} (\mathcal{B}^T\mathcal{B} - \theta^{mT}\theta^m)^2$$

where

$$\mathcal{B} = (\mathbf{W}^T\mathbf{W} + \tilde{\lambda}I)^{-1}\mathbf{W}^T\mathbf{y}$$

One may then replace θ^m with \mathcal{B} . Doing so ensures that the sum of the squared parameters at the top level of the network remains unchanged, but ensures that the (top level of the) penalized loss function reaches its minimum subject to that constraint. This facilitates approximate inference, discussed in section 1.3.

1.3 Inference

Bootstrap approaches to inference will generally be computationally prohibitive, but approximate inference via linear Taylor expansion is feasible[19]. One first computes the estimate of the Jacobian ξ , either numerically or from analytical derivatives. For example, a neural net with two hidden

layers has derivatives

$$\begin{aligned}\frac{\partial \hat{y}}{\partial \beta} &= \mathbf{X} \\ \frac{\partial \hat{y}}{\partial \Gamma_{p1}^1} &= \mathbf{V}_{p1}^1 \\ \frac{\partial \hat{y}}{\partial \Gamma_{p2}^2} &= \sum_{p1} a'(\mathbf{V}^2 \Gamma^2) \Gamma_{p1}^1 \circ \mathbf{V}_{p2}^2 \\ \frac{\partial \hat{y}}{\partial \Gamma_{p3}^3} &= a'(\mathbf{V}^2 \Gamma^2) \Gamma^1 \circ \sum_{p2} a'(\mathbf{Z} \Gamma^3) \Gamma_{p2}^2 \circ \mathbf{Z}_{p3}\end{aligned}$$

where $a'()$ is the derivative of the activation function and the sums are row-wise sums of concatenated column vectors.

Given the Jacobian $\hat{\boldsymbol{\xi}} \equiv \left[\frac{\partial \hat{y}}{\partial \beta}, \frac{\partial \hat{y}}{\partial \Gamma^1}, \frac{\partial \hat{y}}{\partial \Gamma^2}, \frac{\partial \hat{y}}{\partial \Gamma^3} \right]$, one may use it in a manner analogous to a data matrix in an OLS regression to compute the parameter covariance matrix. Assuming homoskedasticity, this would be $\frac{\hat{\boldsymbol{\epsilon}}^T \hat{\boldsymbol{\epsilon}}}{N - p_\xi} \left(\hat{\boldsymbol{\xi}}^T \hat{\boldsymbol{\xi}} \right)^{-1}$. In panel-data settings, the ‘‘cluster-robust’’ estimator[20] may be preferred:

$$\hat{\mathbf{V}}_\theta = \frac{G}{G-1} \frac{N}{N-p_\xi} \left(\hat{\boldsymbol{\xi}}^T \hat{\boldsymbol{\xi}} + \lambda \mathbf{I}^* \right)^{-1} \sum_g \left(\hat{\boldsymbol{\xi}}_g^T \hat{\boldsymbol{\epsilon}}_g^2 \hat{\boldsymbol{\epsilon}}_g^{2T} \hat{\boldsymbol{\xi}}_g \right) \left(\hat{\boldsymbol{\xi}}^T \hat{\boldsymbol{\xi}} + \lambda \mathbf{I}^* \right)^{-1}$$

where G is the number of clusters. If causal inference is desired on parametric terms, $\mathbf{I}^* = \mathbf{I}$ except for those diagonal entries corresponding to the parametric terms, which should be set to zero.⁵

Naturally, penalization induces bias when viewed from a frequentist perspective. Given however that we are not interested in performing inference on penalized elements of $\hat{\boldsymbol{\theta}}$ itself, but rather on functions of $\hat{\boldsymbol{\theta}}$ (i.e.: predictions from the fitted model), it is convenient to adopt the bayesian interpretation of the smoothing process as applied to the linear taylor expansion considered here. Following Wahba [21], Silverman [22], Nychka [23], Ruppert et al. [24], λ can be regarded as the variance on a gaussian prior over the size of $\boldsymbol{\theta}$. If the estimated optimal penalty parameter $\hat{\lambda} \approx \lambda$, Wahba [21] found for smoothing splines that bayesian credible bands around some estimated function $\hat{g}(X)$ derived from the penalized spline estimator have good frequentist properties in an ‘‘across-the-function’’ sense – the true function was found within the estimated credible band for $\hat{g}(X)$, approximately at the nominal rate in large samples. While the individual parameters themselves are biased, this perspective shifts the focus of inference from individual parameters to the smooth functions that they represent. Given the similarities between splines and neural networks – both are overspecified nonlinear transformations of a feature vector⁶ – we proceed from the supposition that this finding may generalize to our case, and explore whether or not it does by

⁵Note that unbiased estimates of marginal effects of parametric terms are only guaranteed when such terms are either exogenous, or when controlling for similarly unpenalized covariates renders them conditionally independent. If $\mathbb{E}[X^T \epsilon | Z] = \mathbf{0}$ and Z enters the model in some form that is subject to penalization, then such controls will not generally solve the endogeneity problem.

⁶While splines transform a feature vector into a form that can be estimated by a (generalized) linear model, we’re concerned here with the linear approximation to a nonlinear model.

simulation in section 2.

The standard errors of the parametric terms may thus be calculated in the usual way from the relevant diagonals of $\hat{\mathbf{V}}_{\theta}$, while the variance of predictions requires the computation of $\tilde{\xi}(\mathbf{Z}^*)$ for new data \mathbf{Z}^* . Pointwise variances are the diagonal of $\tilde{\xi}\hat{\mathbf{V}}_{\theta}\tilde{\xi}^T$.

2 Simulations

This section explores finite-sample properties of the proposed estimator through application to a synthetic dataset, and comparison against standard fixed-effects regression techniques. The synthetic dataset is designed to be intrinsically nonlinear and interactive – a setting in which neural nets should greatly outperform linear models.

2.1 Data-generating process

We generate n_t observations of data \mathbf{Z} for each of n_i cross-sectional units. The data \mathbf{Z}_{it} is distributed multivariate-normally in 5 dimensions, with each cross-sectional unit drawing from its own random mean vector and covariance matrix. The deterministic outcome y_{it}^* is set equal to $\alpha_i + t + \log(\phi(\mathbf{Z}_{it}))$, where ϕ is the standard normal density function in 5 dimensions, α is a fixed effect equal to i , and t is time. As such, the effect of \mathbf{Z} is nonlinear, the marginal effect of $z \in \mathbf{Z}$ is dependent on \mathbf{Z}_{-z} , and is correlated with the group-level intercept. We add noise by setting $y_{it} = y_{it}^* + \mathcal{N}(0, 400)$.

The data are split into training, testing, and validation sets, by time period. Observations in the most “recent” decile comprise the validation set, which is held aside to gauge out-of-sample performance on the selected model. The training and testing sets are divided among the remainder from even and odd-numbered time periods.

2.2 Fitting

These data were fit with panel neural nets of the form $y_{it} = \alpha_i + Pt + f(\mathbf{Z}_{it}) + \epsilon$, as well as with standard fixed-effects models of the form $y_{it} = \alpha_i + Pt + \mathbf{Z}_{it}\boldsymbol{\beta} + \epsilon$. In the neural net, the coefficient on the time trend is left unpenalized in order to observe any bias, and to observe properties of their estimated confidence intervals. The $f(\mathbf{Z}_{it})$ are represented with a 4-layer network, with 12, 11, 10, and 9 nodes from the bottom to the top. The activation function $a(\cdot)$ is the “leaky” rectified linear unit [14];

$$f(x) = \begin{cases} x/100 & \text{if } x < 0 \\ x & \text{if } x > 0 \end{cases}$$

The penalty parameter λ is chosen iteratively. The simulation commences with a large penalty ($\lambda = 8$), and the parameters are estimated subject to that penalty. At convergence, the model

is saved, and $\lambda_{new} \leftarrow \lambda_{old}/2$. The model is then re-fit, starting from the parameter values of the previous model, with some random jitter added to help avoid becoming trapped in local minima.

Fitting proceeds with progressively smaller values of λ until prediction error on the test set fails to improve for three iterations. The selected model is the one that minimizes $\mathbb{E}(y_{test} - \hat{y}_{test})^2$. Note that k -fold cross-validation is generally preferable to the training set/test set approach taken here, but would require each model to be refit k times, and thus would be computationally-prohibitive in our monte-carlo setting.

This process is repeated 1000 times for different draws of the DGP. The selected model is assessed for mean squared error in the validation set, bias in the parametric coefficient, coverage of estimated intervals around the parametric coefficient, and average coverage of estimated intervals around individual predictions in the validation set.

2.3 Results

Results are presented in table 1. The panel neural nets provide substantially better accuracy than do fixed-effects regressions, and provide parametric estimates that are approximately unbiased. Confidence and prediction intervals – on parametric estimates and predictions respectively – cover at approximately nominal rates.

Table 1: Results of monte carlo simulations described in section 2. Standard errors in parenthesis.

T	N. Train	N. Test	N. Val	MSE	MSE (FE)	Bias $\hat{\beta}$	$\hat{\beta}$ coverage	\hat{y} coverage
20	900	900	200	668 (88.16)	1212 (175.5)	-0.0018 (0.1549)	0.967	0.9361
40	900	900	200	627.1 (84.6)	1165 (179.3)	0.001776 (0.07569)	0.987	0.9498
20	1800	1800	400	667.5 (68.96)	1227 (128.1)	0.003283 (0.1113)	0.958	0.9288
40	1800	1800	400	626 (64.12)	1182 (127.4)	0.0002555 (0.05525)	0.958	0.9429
20	2700	2700	600	664.1 (58.8)	1225 (99.36)	-0.0005533 (0.08988)	0.95	0.9268
40	2700	2700	600	624.6 (57.77)	1191 (111.4)	0.00001122 (0.2117)	0.953	0.9416

3 Application: predicting county-level agricultural yields from weather

This final section applies the panel neural net to the prediction of agricultural yields from weather data. This problem is relevant for short-term economic forecasts as well as for longer-range climate change impact assessment.

It is also a natural use-case for a semi-parametric panel data model. First, weather data over an entire year is high-dimensional. While there is only one outcome per year in our setting, the explanatory variables come in the form of several weather variables over hundreds of days of the growing season. Models that penalize high-dimensional covariate sets can help avoid overfitting.

Second, additively-separable models like ridge regression or LASSO may not be appropriate. The “law of the minimum” – a generally-accepted concept in ecology and agronomy – states that plant growth is limited by whatever single growth factor is most constrained. We therefore require an inherently nonlinear and interactive representation of the process that maps weather to yields.

Third, machine-learning algorithms that do not have a linear representation – like decision trees and random forests – have no clear way to incorporate unobserved cross-sectional heterogeneity⁷, which can be important – especially in linear fixed-effect models that do not admit time-invariant regressors.

Finally, agricultural yields in developed countries have seen substantial increases in recent years, driven primarily by technological change. This is naturally represented by a parametric trend, which is incompatible with models like random forests that cannot extrapolate beyond the support of the training data.

3.1 Data

County-level corn yield data for Iowa, Illinois, and Missouri is taken from the National Agricultural Statistical Service of the US Department of Agriculture. We use only county-years with at least 5000 acres planted in corn, and only counties with at least 20 such years. We exclude all counties that have more than 20% of their agricultural land under irrigation. We focus on grain corn, rather than silage.

Weather data comes from PRISM [26]. From its native gridded format, daily data is aggregated into areas corresponding to the agricultural area of each county. We possess measurements of minimum and maximum temperature, precipitation, and dewpoint. These data are subset into a growing season that runs from April through October, leaving us with a total of 856 weather variables.

⁷One exception is the RE-EM trees of Sela and Simonoff [25], which can incorporate random effects into decision tree estimation.

In addition, we augment the weather data with (time-invariant) soils data from SSURGO [27]. The subset we use contains measurements of 38 chemical and physical soil properties, averaged over the county. While such data in linear fixed-effects models will generally be collinear with the individual effects and thereby dropped, the nonlinear nature of the panel neural net allows these variables to moderate the effects of the time-varying weather data. We also include latitude and longitude as regressors, which allows estimated relationships to vary smoothly in space.

3.2 Models

We model these data from the period 1981 to 2014, saving 2015 as a test set for evaluation of our fitted models. These are:

3.2.1 FE-OLS

A standard fixed effects model, we fit

$$y_{it} = \alpha_i + [\text{State}_i \times [\text{time}_t, \text{time}_t^2]] \beta + \mathbf{Z}_{it} \mathbf{\Gamma} + \epsilon_{it}$$

Note that time-invariant soil and location data is collinear with the fixed effects, and dropped from \mathbf{Z} .

3.2.2 LASSO

As above, but solving for $\theta = [\beta, \mathbf{\Gamma}]$ to minimize $\Sigma(y - \hat{y}) + \lambda \Sigma|\theta|$.

3.2.3 Random Forest

The model is

$$y_{it} = g([\text{State}_i \times [\text{time}_t]], \mathbf{Z}_{it}) + \epsilon_{it}$$

and as such does not incorporate any information on the longitudinal nature of the data, except the time trends. We grow 500 trees, with each split in each tree selecting from 1/3rd of the available variables.

3.2.4 Panel neural network

The models is

$$y_{it} = \alpha_i + [\text{State}_i \times [\text{time}_t, \text{time}_t^2]] \beta + g(\mathbf{Z}_{it}) + \epsilon_{it}$$

The quadratic state-by-year time trends are subject to the weight decay penalty, along with $\mathbf{\Gamma}$. We use an architecture of 40 nodes on the bottom layer, 20 in a middle layer, and 10 in a top layer.

The activation function is the rectified linear unit, and training is by batch gradient descent using the adaptive step size algorithm (“RMSprop”) of Tieleman and Hinton [28].

3.2.5 Results

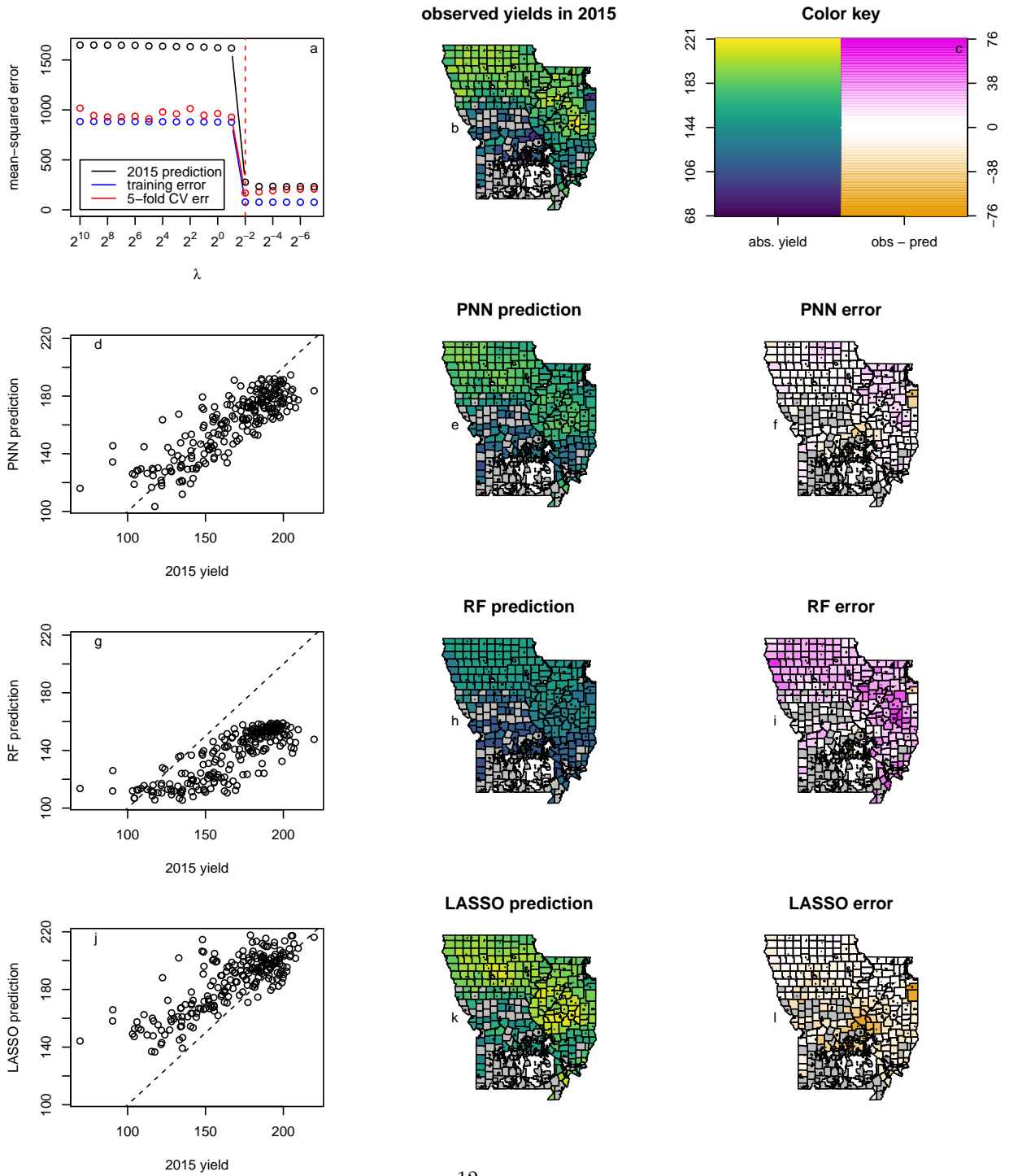
Results are presented in figure 2. Figure 2a shows in-sample, cross-validation, and 2015 prediction errors for the panel neural net – 5-fold cross-validation predicted the 2nd best value of λ out of those tested. For this model, mean-squared error was 273, and predictions were approximately unbiased (figure 2d), though some under-prediction is apparent in north-central Illinois (2f).

The random forest’s MSE was 1197, and its poor performance appears to owe to substantial bias (2g). Indeed, its rampant under-prediction appears to stem from its inability to model a linear time trend such that it can effectively extrapolate one year beyond its training sample – operating like a nearest-neighbor algorithm, random forest will model 2015 as being similar to 2014 and adjacent years, and have no way of discerning a long-term trend in a noisy signal.

The LASSO performs better here than the random forest, with a MSE of 609. Its predictions are biased slightly upwards. This likely stems from over-reliance on the time trend, as this model was unable to discern nonlinear functions the daily training data.

The fixed-effects model is not shown, as the optimal value of λ for the lasso was 0.015 – quite close to zero and thus to the OLS solution.

Figure 2: Prediction results



4 Conclusions and Future Work

This paper has proposed an adaptation of neural networks to panel data, and demonstrated (1) unbiasedness of parametric estimates, (2) good properties of estimated confidence intervals, and (3) efficiency both in a simulated dataset and in an application to yield prediction from weather data.

Further work will provide refinements to the `panelNNET` software used, and demonstrate the applicability of this class of models to the problem of heterogeneous treatment effect estimation from randomized and quasi-experiments.

References

- [1] Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.
- [2] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.
- [3] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
- [4] Stefan Wager, Trevor Hastie, and Bradley Efron. Confidence intervals for random forests: the jackknife and the infinitesimal jackknife. *Journal of Machine Learning Research*, 15(1):1625–1651, 2014.
- [5] Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *arXiv preprint arXiv:1510.04342*, 2015.
- [6] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.
- [7] Trevor J Hastie and Robert J Tibshirani. *Generalized additive models*, volume 43. CRC Press, 1990.
- [8] Daniel J Henderson, Raymond J Carroll, and Qi Li. Nonparametric estimation and testing of fixed effects panel data models. *Journal of Econometrics*, 144(1):257–275, 2008.
- [9] Stefan Hoderlein and Halbert White. Nonparametric identification in nonseparable panel data models with generalized fixed effects. *Journal of Econometrics*, 168(2):300–314, 2012.
- [10] Cong Li and Zhongwen Liang. Asymptotics for nonparametric and semiparametric fixed effects panel models. *Journal of Econometrics*, 185(2):420–434, 2015.

- [11] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, et al. Double machine learning for treatment and causal parameters. *arXiv preprint arXiv:1608.00060*, 2016.
- [12] Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.
- [13] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 807–814, 2010.
- [14] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectifier nonlinearities improve neural network acoustic models. In *Proc. ICML*, volume 30, 2013.
- [15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.
- [16] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.
- [17] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin, 2001.
- [18] Sebastian Ruder. Gradient descent variants. 2016.
- [19] Isabelle Rivals and Léon Personnaz. Construction of confidence intervals for neural networks based on least squares estimation. *Neural Networks*, 13(4):463–484, 2000.
- [20] A Colin Cameron, Jonah B Gelbach, and Douglas L Miller. Robust inference with multiway clustering. *Journal of Business & Economic Statistics*, 2012.
- [21] Grace Wahba. Bayesian” confidence intervals” for the cross-validated smoothing spline. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 133–150, 1983.
- [22] Bernhard W Silverman. Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–52, 1985.
- [23] Douglas Nychka. Bayesian confidence intervals for smoothing splines. *Journal of the American Statistical Association*, 83(404):1134–1143, 1988.
- [24] David Ruppert, Matt P Wand, and Raymond J Carroll. *Semiparametric regression*. Number 12. Cambridge university press, 2003.

- [25] Rebecca J Sela and Jeffrey S Simonoff. Re-em trees: a data mining approach for longitudinal and clustered data. *Machine learning*, 86(2):169–207, 2012.
- [26] Edmund M. Hart and Kendon Bell. *prism: Download data from the Oregon prism project*, 2015. URL <http://github.com/ropensci/prism>. R package version 0.0.6.
- [27] United States Department of Agriculture. Soil Survey Staff, Natural Resources Conservation Service. *Soil Survey Geographic (SSURGO) Database.*, 2016. URL <https://sdmdataaccess.sc.egov.usda.gov>.
- [28] Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2), 2012.
- [29] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
- [30] Wolfram Schlenker and Michael J Roberts. Nonlinear temperature effects indicate severe damages to us crop yields under climate change. *Proceedings of the National Academy of sciences*, 106(37):15594–15598, 2009.
- [31] Yoshua Bengio, Pascal Lamblin, Dan Popovici, Hugo Larochelle, et al. Greedy layer-wise training of deep networks. *Advances in neural information processing systems*, 19:153, 2007.