Снижение размерности пространства зависимой переменной в задачах прогнозирования

Мария Владимирова

Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра интеллектуальных систем

Научный руководитель: д.ф.-м.н. В. В. Стрижов

29 июня 2017 г.

Цели и задачи

Исследуются

Способы обнаружения зависимостей в прогнозируемой переменной.

Проблема

Для построения длительного прогноза требуется последовательно применить прогностическую модель.

Задачи исследования:

- построить алгоритм прознозирования, снижающего мультикоррелированность в пространстве зависимой переменной,
- учесть зависимость между прогностическими переменными, упростив модель и повысив точность прогноза,
- сравнить модели без учета зависимостей между целевыми переменными с учетом линейной, криволинейной и нелинейной зависимости.

Литература

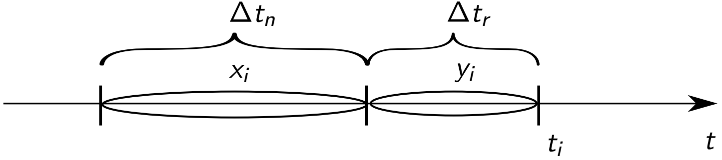
- Höskuldsson A. PLS regression methods //Journal of chemometrics. – 1988. – T. 2. – №. 3. – C. 211-228.
- Frank I. E. A nonlinear PLS model //Chemometrics and intelligent laboratory systems. – 1990. – T. 8. – №. 2. – C. 109-119.
- Rosipal R. Nonlinear partial least squares: An overview //Chemoinformatics and advanced machine learning perspectives. – 2010. – C. 169-189.

Дано:

Временной ряд $\mathbf{x} = [x_t]$, $t = 1, \ldots, n$, $x_t \in \mathbb{R}$.

Требуется спрогнозировать следующие r значений сигнала:

$\mathbf{y} = [y_t]$, $t = 1, \ldots, r$, $y_t \in \mathbb{R}$.



Матрица плана: $\mathbf{X} = [\chi_1, \ldots, \chi_n]$

Матрица ответов: $\mathbf{Y} \in \mathbb{R}^n$

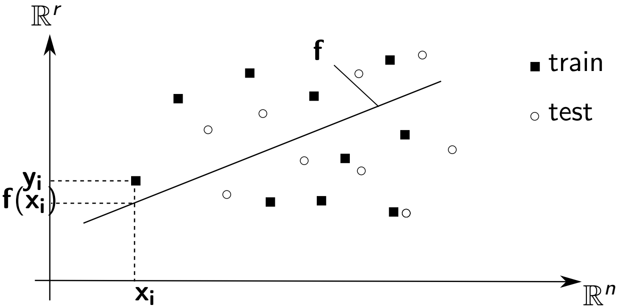
Выборка: $\mathfrak{D} = (\mathbf{X}, \mathbf{Y})$

$$[\mathbf{X} \,|\, \mathbf{Y}] = \left[\begin{array}{ccc|ccc} x_1 & \ldots & x_n & y_1 & \ldots & y_r \\ x_2 & \ldots & x_{n+1} & y_2 & \ldots & y_{r+1} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ x_m & \ldots & x_{m+n-1} & y_m & \ldots & y_{r+m-1} \end{array}\right] = \left[\begin{array}{c|c} \mathbf{x}_1 & \mathbf{y}_1 \\ \mathbf{x}_2 & \mathbf{y}_2 \\ \vdots & \vdots \\ \mathbf{x}_m & \mathbf{y}_m \end{array}\right].$$

Авторегрессионная модель

$$\mathbf{Y} = \mathbf{f}(\mathbf{X}, \boldsymbol{\Theta}) + \varepsilon(\mathbf{X}),$$

где $\mathbf{f}(\mathbf{X}, \boldsymbol{\Theta}) = \mathbf{X}\boldsymbol{\Theta}$ – модель.



$$S(\boldsymbol{\Theta}|\mathfrak{D}) = \left\|\mathbf{f}(\mathbf{X}, \boldsymbol{\Theta}) - \mathbf{Y}\right\|_2^2 = \left\|\mathbf{X}\boldsymbol{\Theta} - \mathbf{Y}\right\|_2^2.$$

$$\boldsymbol{\Theta}^* = \arg\min_{\boldsymbol{\Theta} \in \mathbb{R}^{n \times r}} S(\boldsymbol{\Theta}|\mathfrak{D}).$$

Метод PLSR

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E} \qquad \mathbf{Y} = \mathbf{U}\mathbf{Q}^T + \mathbf{F}$$

$$\mathbf{X}\mathbf{P} \neq \mathbf{T} \qquad \mathbf{Y}\mathbf{Q} \neq \mathbf{U}$$

$$\mathbf{w} \qquad \mathbf{c}$$

$$\mathbf{U} = \mathbf{T}\mathbf{D} + \mathbf{Z}$$

$$\mathbf{X} \in \mathbb{R}^{n\times m},\ \mathbf{T} \in \mathbb{R}^{n\times p}$$
$$\mathbf{Y} \in \mathbb{R}^{n\times r}, \mathbf{U} \in \mathbb{R}^{n\times r}$$

Используется итеративный алгоритм, инициализиируем $\mathbf{u} := \mathbf{y}_1$

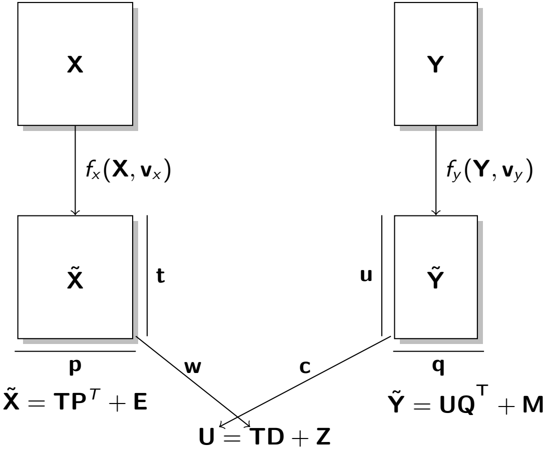
1. $\mathbf{w} = \mathbf{X}^{\mathsf{T}}\mathbf{u}/\|\mathbf{X}^{\mathsf{T}}\mathbf{u}\|$
2. $\mathbf{t} = \mathbf{X}\mathbf{w}$
3. $\mathbf{c} = \mathbf{Y}^{\mathsf{T}}\mathbf{t}/\|\mathbf{Y}^{\mathsf{T}}\mathbf{t}\|$
4. $\mathbf{u} = \mathbf{Y}\mathbf{c}$

$$\mathbf{p} = \mathbf{X}^{\mathbf{T}}\mathbf{t}/\|\mathbf{X}^{\mathbf{T}}\mathbf{t}\|, \quad \mathbf{q} = \mathbf{Y}^{\mathsf{T}}\mathbf{u}/\|\mathbf{Y}^{\mathsf{T}}\mathbf{u}\|$$

Полученное значение $\mathbf{p}$ максимизирует ковариацию между новыми признаками и ответом

$$[Cov(\mathbf{t}, \mathbf{u})]^2 = [Cov(\mathbf{X}\mathbf{w}, \mathbf{Y}\mathbf{c})]^2$$

Мультиколлинеарность в $\mathbf{X}$ и $\mathbf{Y}$



Преобразование $\mathbf{Y}$

- Криволинейное преобразование целевой переменной с вектором параметров $\mathbf{v}$

$$\check{\mathbf{Y}} = g(\mathbf{Y}, \mathbf{v}).$$

Настройка параметров $\mathbf{v}$ с помощью градиентного спуска.

- Нелинейное преобразование целевой переменной с помощью непараметрической функции

$$\tilde{\mathbf{Y}} = h(\check{\mathbf{Y}}) = h(g(\mathbf{Y})) = f(\mathbf{Y})$$

Схема композиции преобразований

$$\mathbf{Y} \xrightarrow{g(\mathbf{Y},\mathbf{v})} \check{\mathbf{Y}} \xrightarrow{h(\check{\mathbf{Y}})} \tilde{\mathbf{Y}} \quad \Rightarrow \quad \mathbf{Y} \xrightarrow{f(\mathbf{Y},\mathbf{v})} \tilde{\mathbf{Y}}$$

Преобразование $\mathbf{X}$

- Криволинейное преобразование зависимой переменной с вектором параметров $\mathbf{v}$

$$\check{\mathbf{X}} = g(\mathbf{X}, \mathbf{v}).$$

- Нелинейное преобразование зависимой переменной с помощью непараметрической функции

$$\tilde{\mathbf{X}} = h(\check{\mathbf{X}}) = h(g(\mathbf{X})) = f(\mathbf{X}).$$

Схема композиции преобразований

$$\mathbf{X} \xrightarrow{g(\mathbf{X},\mathbf{v})} \check{\mathbf{X}} \xrightarrow{h(\check{\mathbf{X}})} \tilde{\mathbf{X}} \quad \Rightarrow \quad \mathbf{X} \xrightarrow{f(\mathbf{X},\mathbf{v})} \tilde{\mathbf{X}}$$

Примеры преобразований

№	Функция	Параметры
1	$g(x) = \mathrm{sign}(x)\exp(a)(\exp(b\|x\|) - 1)$	$a, b > 0$
2	$g(x) = \mathrm{sign}(x)\exp(a)(\exp(b\ln(1 + \|x\|) - 1)$	$a, b > 0$
3	$g(x) = \mathrm{sign}(x)\exp(a)(\exp(b\|x\|^{1/2}) - 1)$	$a, b > 0$
4	$g(x) = \mathrm{sign}(X)\exp(a)(\exp(b\|x\|^{1/3}) - 1)$	$a, b > 0$
5	$g(x) = \mathrm{sign}(x)\exp(a)(\exp(b\|x\|^{1/4}) - 1)$	$a, b > 0$
6	$g(x) = \mathrm{sign}(x)\exp(a)(\exp(b\|x\|^{2}) - 1)$	$a, b > 0$

Таблица: Криволинейные преобразования

Алгоритм cnIPLS для зависимой переменной

Криволинейное + нелинейное преобразование f в пространстве целевой переменной $\mathbf{Y}$ в алгоритме PLSR

Инициализировать $\mathbf{v}$, $\mathbf{Y} := \mathbf{y}_1$, $\mathbf{u}_0 := f(\mathbf{Y}, \mathbf{v})$, выполнить в цикле

1. $\mathbf{w} = \mathbf{X}^{\mathsf{T}}\mathbf{u}_0 / \|\mathbf{X}^{\mathsf{T}}\mathbf{u}_0\|$
2. $\mathbf{t} = \mathbf{X}\mathbf{w}$
3. $\tilde{\mathbf{Y}} = f(\mathbf{Y}, \mathbf{v})$
4. $\mathbf{c} = \tilde{\mathbf{Y}}^{\mathsf{T}}\mathbf{t} / \|\tilde{\mathbf{Y}}^{\mathsf{T}}\mathbf{t}\|$
5. $\mathbf{u} = \tilde{\mathbf{Y}}\mathbf{c}$
6. $\mathbf{e} = \mathbf{u} - \mathbf{u}_0$
7. $\mathbf{J} = \partial\mathbf{u}/\partial\mathbf{v}$
8. $\Delta\mathbf{v} = (\mathbf{J}^{\mathsf{T}}\mathbf{J})^{-1}\mathbf{J}^{\mathsf{T}}\mathbf{e}$
9. $\mathbf{v} = \mathbf{v} + \Delta\mathbf{v}$, $\|\mathbf{v}\| \to 1$
10. $\mathbf{u}_0 := \mathbf{u}$

Вычислить $\mathbf{p} = \mathbf{X}^{\mathsf{T}}\mathbf{t} / \|\mathbf{X}^{\mathsf{T}}\mathbf{t}\|$, $\mathbf{q} = \tilde{\mathbf{Y}}^{\mathsf{T}}\mathbf{u} / \|\tilde{\mathbf{Y}}^{\mathsf{T}}\mathbf{u}\|$.

Эксперимент

- данные содержат временные ряды электроэнергии, погоду (температура, осадки, сила ветра, влажность, солнечная энергия), график отпусков;
- потребление энергии измерялось ежечасно с 1999 по 2004 год, 52512 наблюдений;
- погода измерялась ежедневно, 2188 наблюдений.

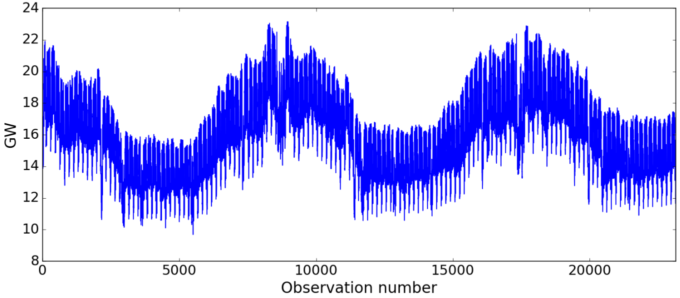
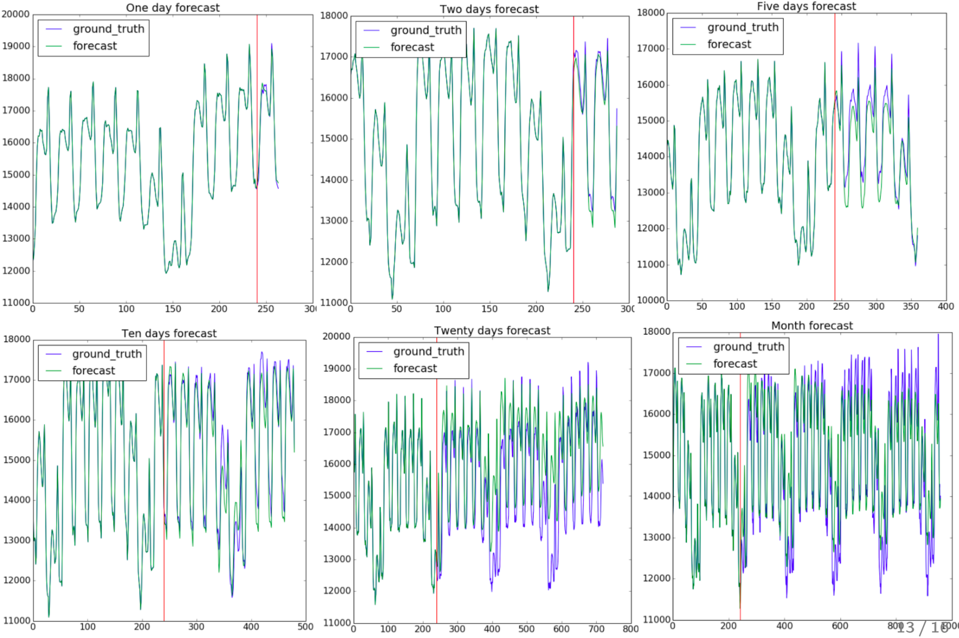
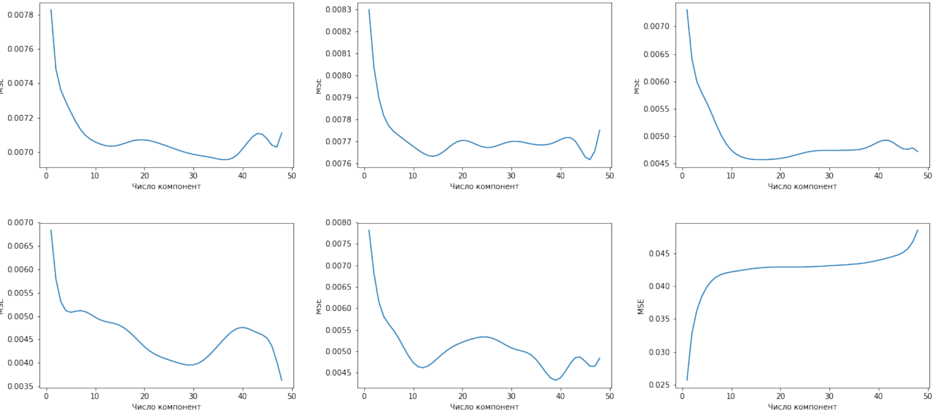


Рисунок: Временной ряд потребления электроэнергии

Предсказания на разные временные отрезки



Зависимость ошибки от числа компонент для разных функций преобразования



Результаты

Алгоритм	**N=3**	**N=5**	**N=10**	**N=20**
PLS	0,00404	0,00337	**0,00151**	0,00135
cnlPLS $g(x) = \text{sign}(x)e^a(\exp(b\|x\|) - 1)$	0.00529	0.00514	0.00536	0.00506
cnlPLS $g(x) = \text{sign}(x)e^a(\exp(b\ln(1 + \|x\|) - 1)$	0.00362	0.00386	0.00326	0.00317
cnlPLS $g(x) = \text{sign}(x)e^a(\exp(b\|x\|^{1/2}) - 1)$	0.00272	0.00236	0.00287	**0.00128**
cnlPLS $g(x) = \text{sign}(x)e^a(\exp(b\|x\|^{1/3}) - 1)$	**0.00241**	**0.00233**	0.00221	0.00173
cnlPLS $g(x) = \text{sign}(x)e^a(\exp(b\|x\|^{1/4}) - 1)$	0.00796	0.00768	0.00737	0.00803
cnlPLS $g(x) = \text{sign}(x)e^a(\exp(b\|x\|^2) - 1)$	0.00816	0.00798	0.00796	0.00775

Таблица: Значения ошибки MSE для разных чисел компонент и разных функций

Результаты, выносимые на защиту

- Предложены алгоритмы прогнозирования временных рядов.
- Предложены методы снижения мультиколлинеарности в пространстрах зависимой и независимой переменной.
- Выполнена программная реализация и проведены численные эксперименты, показавшие повышение качества решения задачи прогнозирования сигналов.