Credibility evaluation of income data with hierarchical correlation reconstruction

Jarosław Duda[⋆] Adam Szulc[†]
[⋆] Jagiellonian University, Cracow, Poland, Email: dudaj@interia.pl
[†] Warsaw School of Economics, Warsaw, Poland

***Abstract*—In situations like tax declarations or analyzes of household budgets we would like to evaluate credibility of exogenous variable (declared income) based on some available (endogenous) variables - we want to build a model and train it on provided data sample to predict (conditional) probability distribution of exogenous variable based on values of endogenous variables. Using Polish household budget survey data there will be discussed simple and systematic adaptation of hierarchical correlation reconstruction (HCR) technique for this purpose, which allows to combine interpretability of statistics with modelling of complex densities like in machine learning. For credibility evaluation we normalize marginal distribution of predicted variable to $\rho \approx 1$ uniform distribution on $[0,1]$ using empirical distribution function ($x = EDF(y) \in [0,1]$), then model density of its conditional distribution (Pr$(x_0|x_1x_2\ldots)$) as a linear combination of orthonormal polynomials using coefficients modelled as linear combinations of properties of the remaining variables. These coefficients can be calculated independently, have similar interpretation as cumulants, additionally allowing to directly reconstruct probability distribution. Values corresponding to high predicted density can be considered as credible, while low density suggests disagreement with statistics of data sample, for example to mark for manual verification a chosen percentage of data points evaluated as the least credible.**

I. Introduction

While in standard regression we want to estimate the conditional expected value, in some situations we need to predict the entire probability distribution. In ARMA/ARCH modelling [1], or data compression like JPEG-LS [2], it is resolved by predicting the most likely value, usually using a linear combination of neighboring values, then assuming some parametric distribution of error from this prediction - for example as Gaussian in ARMA-like models or Laplace distribution in data compression. Widths of such distributions are often chosen based on the context (e.g. in ARCH, JPEG-LS).

However, in situation like credibility evaluation of tax declarations, the conditional distribution of the main variable to verify (declared income) is quite complex, as we will see in the presented analysis, and we need to model dependencies with multiple categorical variables - their properties might affect not only the expected value, but also higher moments. Such dependence between expected values of two variables is described by correlation coefficient, between variances e.g. in ARCH model, HCR used here allows to systematically exploit such dependencies of any moments between two or more variables.

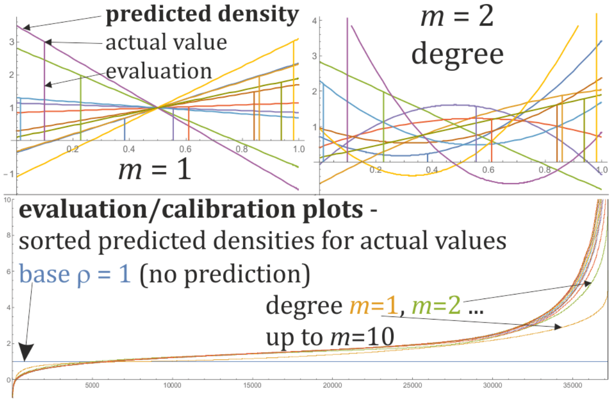

Figure 1. Top: modelled conditional densities of Pr$(x_0|x_1\ldots x_d)$ using degree $m=1$ (left) or $m=2$ (right) polynomials (Fig. 3 contains further up to $m=10$) for predicted exogenous variable (equivalent income normalized to uniform marginal distribution on $[0,1]$) based on the remaining (endogenous) variables for ten randomly chosen data points from the sample. The actual values are marked with vertical lines, the higher predicted density for them, the better - we see that in most cases the prediction is above the base $\rho=1$. Inconvenience of this method is sometimes obtaining negative densities, but we can interpret such predictions, e.g. as just having low credibility here. Bottom: evaluation/calibration plots for predictions with $m=1\ldots10$ degree polynomials, obtained by sorting predicted densities for actual values. We see that in $\approx 1/6$ of cases this prediction is worse than $\rho=1$, but it can be essentially better in the remaining cases, especially if using high degree polynomial. Such low predicted density suggests disagreement of given data point with statistics of data sample - suggests its low credibility, for example to perform manual verification if this value is below a threshold chosen to mark a given percentage of the least credible.

To evaluate credibility of a given (exogenous) variable y_0, we would like to predict its conditional probability density of Pr$(y_0|y_1\ldots y_d)$ based on the remaining (endogenous) variables $y_1\ldots y_d$ - intuitively, the higher such density is, the higher credibility of this value. However, this intuition requires some normalization, as for example tails of distributions have lower density what should not be interpreted as lower credibility - these values are just spread over wider range.

A natural normalization, used for example in copula theory [3], is to nearly uniform marginal distribution $\rho_0(x_0) \approx 1$ on $[0,1]$. We can use empirical distribution function (EDF, by sorting obtained values) to transform the original variable y to $x = EDF(y)$ from nearly uniform distribution. Using this normalization, modelled $\rho_0(x_0)$ density of Pr$(x_0|x_1\ldots x_d)$ can be seen as evaluation of credibility, its examples are presented in Fig. 1.

To model $\Pr(x_0|x_1 \dots x_d)$ we will use hierarchical correlation reconstruction (HCR) approach ([4], [5], [6], [7]): model density of joint distribution of all variables $[0,1]^{d+1}$ as a linear combination of orthonormal polynomials:

$$\rho(\mathbf{x}) = \sum_{\mathbf{j}=(j_0 \dots j_d)} a_\mathbf{j}\, f_{j_0}(x_0) \cdot \ldots \cdot f_{j_d}(x_d) \qquad (1)$$

where $(f_j)_j$ satisfy $\int_0^1 f_i(x) f_j(x) dx = \delta_{ij}$.

Such $a_\mathbf{j}$ coefficients have cumulant-like interpretation. For example $a_{100\dots0}$ has similar behavior as expected value of the first variable, $a_{020\dots0}$ as variance of the second variable. We can also use mixed coefficients, for example $a_{110\dots0}$ describes dependence between expected values of the first two variables - has similar interpretation as correlation coefficient. We can model complex joint density with such polynomial approximation, especially that mean-square estimation of these coefficient turns out very inexpensive and can be calculated independently [4]: using orthonoromal basis, coefficient of a given function turns out just average of this function over the sample: $a_\mathbf{j} = \frac{1}{n}\sum_{1=1}^n f_\mathbf{j}(\mathbf{x}^i)$ for $\mathbf{x}^i = (x_0^i, \dots, x_d^i)$ data points.

For credibility evaluation, after normalizing all marginal distributions to nearly uniform on $[0,1]$, we would like to estimate their joint density as such linear combination - using a chosen basis, for example exploiting only pairwise dependencies (up to two nonzero indexes). Then substituting $x_1 \dots x_d$ coordinates to such joint distribution, and normalizing to integrate to 1, we get probability density $\rho_0(x_0)$, describing credibility of this value.

However, above HCR approach is appropriate for continuous variables, while here we have many discrete - there will be suggested adaptation for this situation. Finally, for simplicity and interpretability there will be just used least-squares regression to optimize linear dependencies between properties of endogenous variables (v) and cumulant-like coefficients of the predicted variable:

$$\rho_0(x_0) = 1 + \sum_j a_j f_j(x_0) \quad \text{for} \quad a_j = \beta_0^j + \sum_k \beta_k^j v_k$$

with coefficients chosen by least-square optimization of $\sum_i \|f_j(x_0^i) - a_j\|^2$.

For evaluation/calibration there will be used plot of sorted $\Pr(x_0^i|x_1^i \dots x_d^i)$ for all n points from the dataset, presented in Fig. 1. Its horizontal axis can be seen as quantile regarding modelled credibility. For example choosing that we want to manually verify 1% of least credible data points, we can read the density threshold ($\rho \approx 0$) from this plot and verify only those below this threshold.

II. Dataset

The dataset is composed of observations on 37215 households collected annually by the Central Statistical Office of Poland (GUS). There will be used the following variables (all but the first are endogenous - they are used to predict the first variable which is exogenous; variable names are provided in brackets):

1) Continuous: equivalent monthly income[1] (inceq) in PLN - the exogenous variable, remaining equivalent cash in PLN at the end of month (casheq), shares of expenditures on luxury goods and food (luxury and food, respectively),
2) Discrete ordinal (the number of distinct values is provided in brackets): age of the household head in years (age, 87 values), his/her completed education level (edu, 11), number of persons (pers, 14), number of younger (child1, 8) and older (child2, 7) children in the household, urbanisation level (urb, 3), type of residence (loc, 6), number of cars (cars, 7), month of the query (month, 12).
3) Discrete categorical: main income source (source1, 12) and additional source (source2, 13) , voivodship (voi, 16), building type (build, 4), its ownership type (own, 6), age of the newest car (carage, 5), subjective evaluations of: change in the material position (chg, 5), income sufficiency (sf, 5), level of satisfaction of needs for food (sff, 5), clothing (sfc, 5), health care (sfh, 6), housing fees (sfs, 6), housing equipment (sfq, 6), culture (sfl, 7), education (sfe, 6), tourism and recreation (sft, 6).
4) Binary: sex of the household head, subjective evaluation whether the dwelling is too small or too large, presence in the household of persons: with tertiary and secondary education, unemployed, handicapped.

The model will use linear combinations of their properties. Age will be treated as continuous variable, their pairwise dependencies can be seen in Fig. 2, their cumulant-like proprieties are used in linear combination for prediction. To avoid arbitrary choice of weights, all the remaining variables are treated as binary: split into 0/1 variables, as many as the number of distinct values, being 1 if the category agrees, 0 otherwise.

III. Normalization and orthognonal polynomial basis

For many reasons it is convenient to normalize variables to have nearly uniform marginal distributions: to see them as quantiles - interpreting ranges as population percentages, to directly interpret predicted density as credibility here, finally to use polynomials for density estimation with normalized coefficients. In previous applications of HCR ([6], [7]), there was used CDF (cumulant distribution function) of approximated parametric distribution (Laplace) for this normalization - approximating general behavior, then modelling with polynomial corrections from this idealization, which can evolve in time for non-stationary time series. Here probability distribution is stationary and too complex for parametric distributions, hence, like in copula theory [3], we will directly use EDF for this normalization.

[1] It is calculated as the household disposable income divided by a respective equivalence scale i. e. an indicator supposed to measure impact of demographic variables on cost of living. In the present study simple OECD 70/50 formula is employed: scale = 1 + (number of adults minus one) · 0.7 + (number of children) · 0.5.

A. Normalization

Having $y^1, \ldots, y^n$ sample, we can normalize it with EDF by sorting the values - finding order (bijection) $o : \{1 \ldots n\} \to \{1 \ldots n\}$ such that:

$$y^{o(1)} \leq y^{o(2)} \leq \ldots \leq y^{o(n)} \tag{2}$$

Hence, y^i is in $o^{-1}(i)$-th position of this order - wanting them to have nearly uniform distribution on $[0,1]$, a natural choice is $x^i = \frac{1}{n}(o^{-1}(i) - 1/2)$.

However, especially for discrete variables, (2) has many equalities, what needs special treatment - there is no base to choose an order among equal values, all of them should be transformed into the same value x^i, naturally chosen as the center of such range - we can see one such horizontal line for casheq = 0 in Fig. 2, and age sample consisting only of horizontal lines due to rounding to complete years.

Finally, the used generalized formula (working for both continuous and discrete variables) is:

$$x^i = \frac{\min\{k : y^{o(k)} = y^i\} + \max\{k : y^{o(k)} = y^i\} - 1}{2} \tag{3}$$

We can use this formula to normalize each variable of $\mathbf{y}^i = (y^i_0, \ldots, y^i_d)$: separately for each lower index, getting $\mathbf{x}^i = (x^i_0, \ldots, x^i_d)$, which for continuous variables have nearly uniform marginal distribution on $[0,1]$. However, for discrete variables it is a step function, especially for binary - needing a completely different treatment.

There are ways to transform discrete variables into a smaller number of nearly independent continuous variables - for example by choosing a set of orthonormal projections. This choice is sometimes done randomly, but it often can be optimized by dimensionality reduction techniques like PCA (principal component analysis) or its discrete analogues like MCA (multiple correspondence analysis) [8]. Such projections often have nearly Gaussian distribution, but we would loose interpretability this way - this article presents analysis directly using discrete variables to get better interpretation of obtained coefficients (Fig. 4), however, it might be worth to explore also methods like MCA to directly work on continuous variables.

B. Orthogonal polynomial basis and HCR

Assuming (normalized) variable is from nearly uniform distributions on $[0,1]$, it is very convenient to represent its density with polynomial: $\rho(x) = \sum_j a_j f_j(x)$. Using orthonormal basis $\int_0^1 f_i(x) f_j(x)\, dx = \delta_{ij}$, mean-square optimization leads to [4] inexpensive estimation by just averaging: $a_j = \frac{1}{n}\sum_{i=1}^n f_j(x^i)$.

The first five of these polynomials (rescaled Legendre) are $f_0 = 1$, and f_1, f_2, f_3, f_4 correspondingly:

$$\sqrt{3}(2x-1), \sqrt{5}(6x^2-6x+1), \sqrt{7}(20x^3-30x^2+12x-1),$$
$$3(70x^4 - 140x^3 + 90x^2 - 20x + 1).$$

For normalization we need $a_0 = 1$. The a_1 term shifts the expected value toward left or right. Positive a_2 term increases probability of extreme values, reducing it

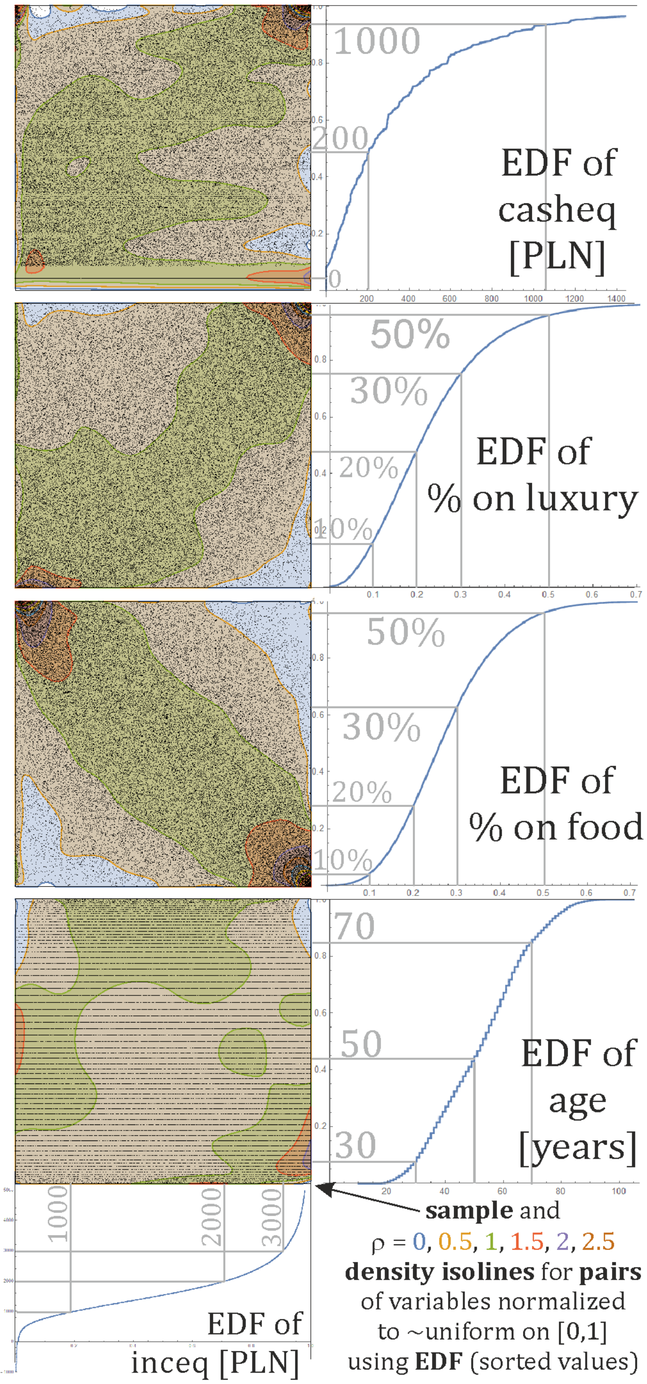

Figure 2. Pairwise dependencies between 5 variables treated as continuous: exogenous (income) on horizontal axis to be predicted from endogenous variables on vertical axis. Each is normalized to nearly uniform marginal distribution - position can be seen as quantile, 0.5 as median, length e.g. 0.2 as 20% of population. Some of them have discreteness - corresponding to horizontal lines. Each of four $[0,1]^2$ shown regions contains 37215 black dots from the dataset, and isolines for their density (would be $\rho \approx 1$ for independent) - estimated with HCR as polynomial $\sum_{ij=0}^9 a_{ij} f_i(x_0) f_j(x_1)$ using 100 coefficients (mixed moments up to 9th). For example for age we can see that younger people have higher expected inceq, middle-age lower, older closer to median - we need at least second order polynomial (f_2) of age to model such behavior.

for central values - has analogous behavior as variance. And so on: a_j coefficient has similar interpretation as j-

th cumulant. Using degree m polynomial: $j = 0\ldots m$ corresponds to modelling distribution using the first m moments, additionally directly getting density estimation from them.

We can also exploit statistical dependencies between two or more variables this way - by analogously modelling joint distribution on $[0,1]^{d+1}$ using product basis: $\rho(\mathbf{x}) = \sum_{\mathbf{j}=j_0\ldots j_d} a_\mathbf{j}\, f_{j_0}(x_0)\cdot\ldots\cdot f_{j_d}(x_d)$. This way $a_\mathbf{j}$ represents mixed cumulants - their dependencies between multiple variables. For a large number of variables, most of coordinates of used $\mathbf{j}$ should be 0 - coefficients with single nonzero coordinate describe probability distribution of corresponding variable, with two nonzero describe pairwise dependencies and so on - getting hierarchical correlation reconstruction (HCR) of a given distribution.

We could directly use such modelled joint density $\rho(x_0\ldots x_d)$ for credibility evaluation ($\Pr(x_0|x_1,\ldots,x_d)$) by just substituting $x_1\ldots x_d$ and normalizing obtained polynomial of x_0 to integrate to 1. This way $f_{j_0}(x_0)$ is expressed as nearly a linear combination of various products of $f_{j_k}(x_k)$. However, this approach has difficulties with discrete values - hence, there is finally used linear regression to directly optimize coefficients of such linear combinations.

Orthogonal polynomial basis allows also for a different perspective: for a given $x \in [0,1]$ value, we can take its coordinates in this basis: $(f_j(x))_j$. As estimated a_j is just average over such corresponding coordinates, we can imagine e.g. $f_1(x)$ as contribution of this point to expected value, $f_2(x)$ to variance etc. Hence, we can use $(f_j(x))_j$ as a list of properties of a given data point to infer its other property, exploiting nonlinear dependencies this way (for $j \geq 2$) - we will use such properties in linear regression for prediction here.

There is generally a large freedom of choice for properties used in inference. For example for age variable here, a standard approach is dividing into age ranges, what can be seen as using $(f_j(x))_j$ with family of functions being 1 on a given age range and 0 otherwise. It leaves a question of sizes of these age ranges: long ranges reduce data precision, short ranges mean lower statistics and not exploiting local behavior (of neighboring ranges). Normalizing variable and using $(f_j(x))_j$ with orthonormal family of polynomials is an attempt to automatically exploit local dependencies, making each coefficient being affected by all data points. Numerical tests for age gave slightly more accurate predictions using orthonormal polynomials, than using the same number of fixed length age ranges.

C. *Handling discrete values*

Analogously to continuous variables with $\langle f,g\rangle = \int_0^1 f(x)g(x)dx$ scalar product, we could directly work on discrete variables with $\langle f,g\rangle_d = \sum_x w_x f_x g_x$ product using some weight w_x - using basis of orthonormal vectors this time, we can we get analogous mean-square estimation as $a_j = \frac{1}{n}\sum_i f_{jx^i}$.

However, while f_0 should be chosen as marginal distribution in analogy to continuous case, choosing the weights is a nontrivial problem, planned to be explored in the future. For simplicity and to present intuition of treating $(f_j(x))_j$ as useful properties for inference, which can be imagined as x-th contribution to j-th moment, in the presented analysis there will be used linear regression to directly optimize coefficients.

IV. Used algorithm

The currently used algorithm optimizes coefficients with least-square regression:

1) All variables treated as continuous - including casheq having a large percentage of exactly 0 value, and age obtaining 87 distinct discrete values - are normalized to nearly uniform marginal distribution using formula (3).
2) All the remaining variables are treated as categorical and transformed into binary - thanks of it, weights of individual categories are optimized in later regression. For example edu(cation) obtains 11 different values, hence it is transformed into 11 binary variables: each being 1 if category agrees, 0 otherwise.
3) Denote $v(y_1\ldots y_d)$ as vector built of all properties of endogenous variables (normalized or not) directly used for prediction as a linear combination - here it has 223 coordinates visualized in Fig. 4. Its zeroth coordinate is fixed as 1 to get constant term (β_0) in later regression. Then for variables treated as continuous, it contains $f_j(x_k)$ for $j = 1$ up to a chosen degree, which in 9 here - for all 4 variables treated this way (casheq, luxury, food, age) - getting $4 \times 9 = 36$ coordinates of v. It further contains all the remaining variables - categorical transformed into binary, and the original binary variables - both use only 0 or 1 values.
4) Build e.g. $n\times 223$ matrix M with rows being applied $v(y_1^i\ldots y_d^i)$ function to all $i = 1\ldots n$ data points.
5) For each property we would like to use least square linear regression to infer $f_j(x_0)$ for $j = 1\ldots m$ to predict density as degree m polynomial. For this purpose, build vectors $b^j = (f_j(x_0^i))_{i=1\ldots n}$, then find vectors β^j minimizing $\|M\beta^j - b^j\|$. It can be realized with pseudoinverse, and is implemented in many numerical libraries, e.g. as "LeastSquares[M,b]" in Wolfram Mathematica. Values of these final used coefficients are visualized in Fig. 4.
6) Now predicted density is

$$\rho_0(x_0) = 1 + \sum_{j=1}^{m} a_j f_j(x_0) \quad \text{for} \quad a_j = v(y_1\ldots y_d)\cdot\beta^j$$

This predicted density is for exogenous variable normalized to nearly uniform marginal distribution on $[0,1]$ - what allows to use it directly for credibility evaluation.

Sorting such predicted densities of actual values, we get evaluation/calibration plot presented in Fig. 1 for various used degree $m = 1 \ldots 10$. Inconvenience of parameterizing density as polynomial is sometimes obtaining negative value, what need proper interpretation. It happens frequently in predictions in Fig. 3, but only in $\approx 1\%$ of cases in this plot of sorted predictions - confirming prediction and allowing to exclude such negative density regions with near certainty. Predicted negative density suggests disagreement of this point with sample statistics - low credibility of this data point.

To apply this evaluation/calibration plot, we can rescale its horizontal axis to $[0, 1]$ range, allowing to interpret it as quantile in credibility evaluation. Then for example choosing a percentage of least credible data points for manual verification, we can read density threshold for this percentage, then mark for further verification all data points below this threshold (e.g. $\rho_0(x_0) < 0$ for $\approx 1\%$).

V. Conclusions and further work

There was presented a simple general approach for applying HCR technique with discrete variables - on example of credibility evaluation for income data, allowing to model conditional probability distribution for (normalized) predicted variable.

It tests agreement with statistics of provided data sample - will not detect systematic improper behavior existing in this sample. Handling such situations would rather require some supervised learning, like manually marking some of suspicious data points and then evaluate probability of being in this marked set.

The main purpose of this article was methodological - presenting simple interpretable analysis, but leaving many possibilities for improvements.

For example it has exploited only pairwise dependencies - between the exogenous variable and (separately) each of endogenous variable. We can analogously include higher order dependencies by using products of considered variables in vector v of above algorithm, e.g. include $f_{j_1}(x_1)f_{j_2}(x_2)$ properties in the least square regression for 3-point correlation, and analogously for higher order dependencies. As their number grows exponentially with dependency order, this choice requires some selectivity - for example can be done in hierarchical way: e.g. search for non-negligible 3-point correlations by expanding essential 2-point correlations. In presented analysis the number of parameters is negligibly small comparing do sample size - there is no danger of overfitting. Otherwise, the sample should be split at least into training and evaluation set [7].

Also, the main purpose of using least-square regression was simplicity and interpretability - it probably can be improved, what is planned to be explored in the future.

As mentioned, another direction worth exploring is using dimensionality reduction techniques like PCA or its discrete analogue: MCA, to reduce the number of variables and make them continuous.

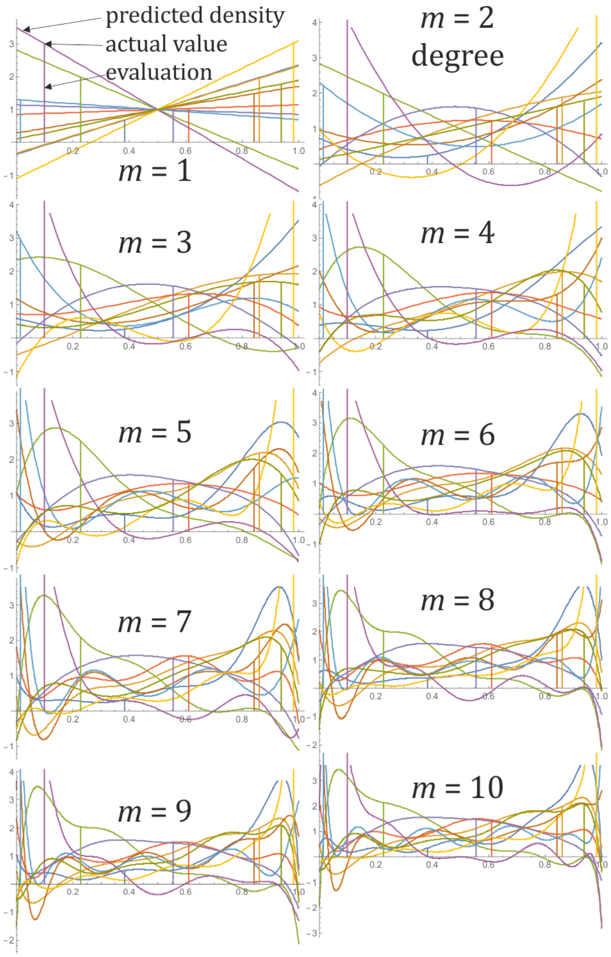

Figure 3. Predicted densities of $\Pr(x_0|x_1 \ldots x_d)$ like in Fig. 1, but for all $m = 1, \ldots, 10$ polynomial degrees, what can be imagined as modelling up to m-th moment. All use the same randomly chosen 10 data points - focusing on a chosen color, we can follow evolution of its prediction with growing degree. For example the extreme ones (violet and yellow) have essentially improved while using parabola, but the remaining stay nearly unchanged. We can see that localization of predicted range usually improves with degree.

References

[1] T. C. Mills and T. C. Mills, *Time series techniques for economists*. Cambridge University Press, 1991.

[2] M. J. Weinberger, G. Seroussi, and G. Sapiro, "The loco-i lossless image compression algorithm: Principles and standardization into jpeg-ls," *IEEE Transactions on Image processing*, vol. 9, no. 8, pp. 1309–1324, 2000.

[3] F. Durante and C. Sempi, "Copula theory: an introduction," in *Copula theory and its applications*. Springer, 2010, pp. 3–31.

[4] J. Duda, "Rapid parametric density estimation," *arXiv preprint arXiv:1702.02144*, 2017.

[5] ——, "Hierarchical correlation reconstruction with missing data," *arXiv preprint arXiv:1804.06218*, 2018.

[6] ——, "Exploiting statistical dependencies of time series with hierarchical correlation reconstruction," *arXiv preprint arXiv:1807.04119*, 2018.

[7] J. Duda and M. Snarska, "Modeling joint probability distribution of yield curve parameters," *arXiv preprint arXiv:1807.11743*, 2018.

[8] M. Greenacre and J. Blasius, *Multiple correspondence analysis and related methods*. Chapman and Hall/CRC, 2006.

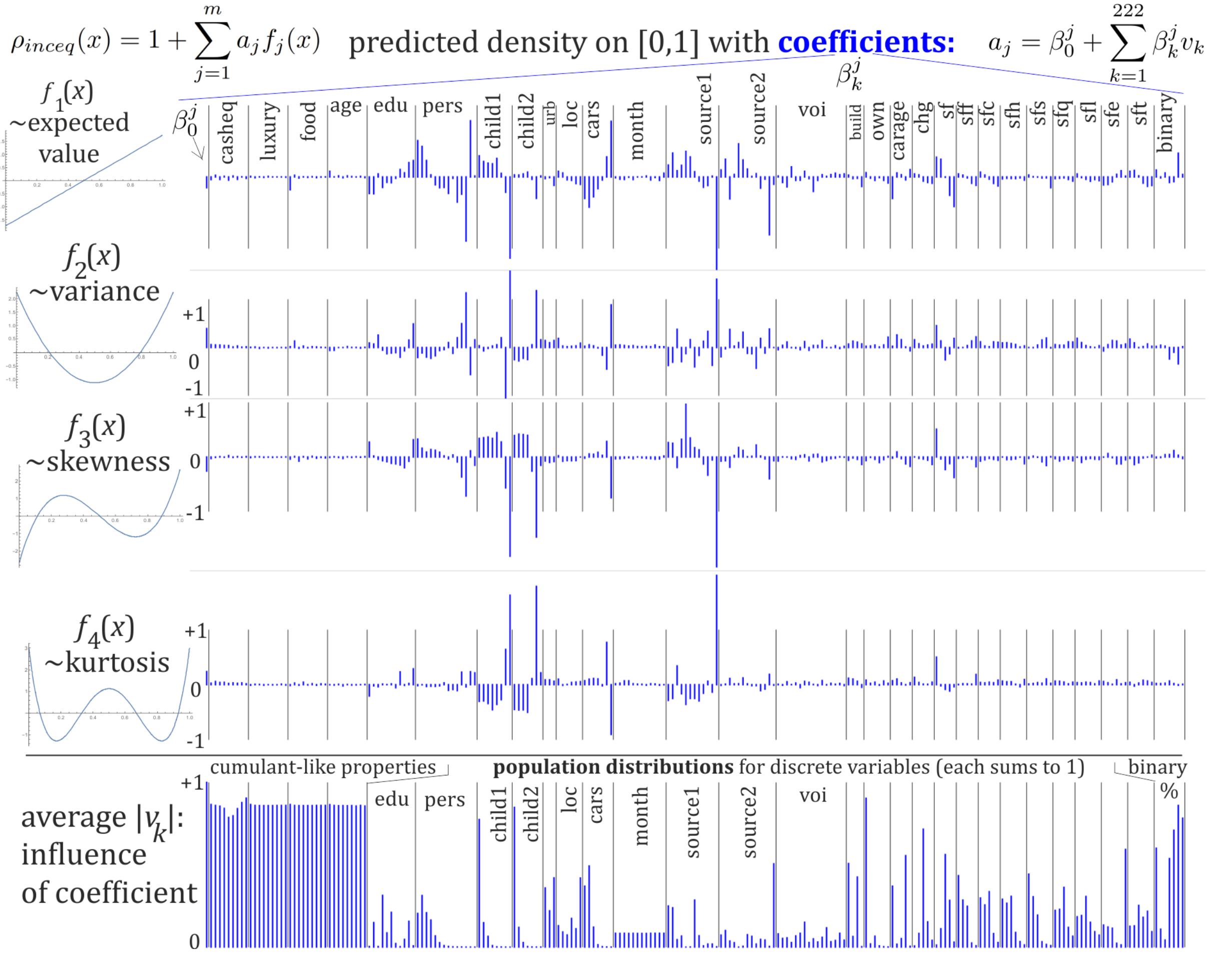

Figure 4. Top: final 4×223 coefficients obtained by least-square regression (optimizing $\sum_i \|f_j(x_0^i) - a_j\|^2$) for predicting probability density of exogenous variable (equivalent income) based on properties of endogenous variables - independently for coefficients corresponding to expected value (f_1), variance (f_2), skewness (f_3) and kurtosis (f_4) of predicted variable. For endogenous variables treated as continuous (casheq, luxury, food, age), the used properties are $f_j(x_l)$ for $j = 1 \ldots 9$ describing j-th cumulant-like behavior. The remaining variables are treated as binary variables (0 or 1) for each appearing possibility. For example in the first row, negative first coefficient for food connects their expected values - describes anticorrelation, analogously for age it is positive - they are correlated. Coefficients for voivodeships (voi) describe individual corrections for each geographical region. Low statistics for some coefficients can lead to surprising behavior, for example we can see reduction of equivalent income with the number of persons in the household (pers), with a surprising spike at the end - it corresponds to a single data point of 12 person household. Bottom: average $|v_k|$ describing average influence of β_k coefficient - it is fixed 1 for v_0, nearly constant for cumulant-like coefficients for continuous variables, for each binarized discrete variable its contributions sum to 1: they describe proportions in population, for binary variables it is in $[0, 1]$.