Full conditional distributions for handling restrictions in the context of automated statistical data editing

United Nations Economic Commission for Europe – Conference of European Statisticians
Expert meeting on Statistical Data Editing
7-9 October 2024, Vienna
Christian Aßmann, Ariane Würbach, Florian Dumpert, Younes Saidani

LIfBi
LEIBNIZ-INSTITUT FÜR BILDUNGSVERLÄUFE

Problem description

- The adjustment of implausible values in a data set $X = (X_1, \ldots, X_P) = (X_1^{\text{ipl}}, X_1^{\text{obs}}, \ldots, X_P^{\text{ipl}}, X_P^{\text{obs}}) = (X_{\text{ipl}}, X_{\text{obs}})$ is phenomenologically very similar to the handling of missing values in statistical analyses.
- Replace the implausible values using suitable estimation functions based on the density function
$$f(X_{\text{ipl}}|X_{\text{obs}}).$$
- Statistical modelling, including the methods of statistical machine learning, approximate the conditional distribution
$$f(X_p|X_{\setminus p}, \theta_p).$$
- The actual target density of interest is therefore not directly accessible.

Sequentielle Approximation von $f(X_{\text{ipl}}|X_{\text{obs}})$

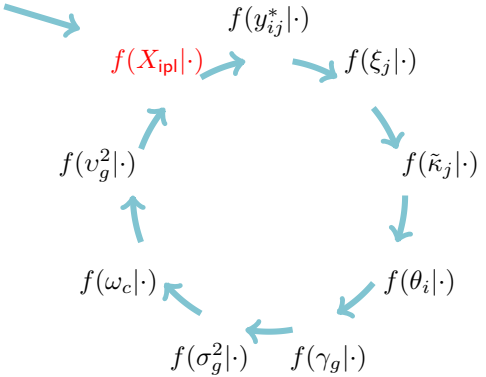
- The approximation of $f(X_{\text{ipl}}|X_{\text{obs}})$ is performed using the extended density $f(X_{\text{ipl}}, \theta|X_{\text{obs}}))$.
- The extended density can be obtained by iterating through the two fully conditional densities $f(\theta|X_{\text{ipl}}, X_{\text{obs}})$ and $f(X_{\text{ipl}}|\theta, X_{\text{obs}})$.
- As a rule, subsequences must be considered for the individual components, e.g.

$$f(\theta|X_{\text{ipl}}, X_{\text{obs}}) = \prod_{p=1}^{P} f(\theta_p|X_{\text{ipl}}, X_{\text{obs}}) \quad \text{or} \quad f(X_{\text{ipl}}|\theta, X_{\text{obs}}) = \prod_{p=1}^{P} f(X_p^{\text{ipl}}|\theta_p, X_{\setminus p}^{\text{ipl}}, X_{\text{obs}}),$$

  which can be derived from $\prod_{p=1}^{P} f(X_p|X_{\setminus p}, \theta_p) f(\theta_p)$.
- The central component of the approximation is the use of *data augmentation*, i.e. $\prod_{p=1}^{P} f(X_p^{\text{obs}}, X_p^{\text{ipl}}|X_{\setminus p}^{\text{obs}}, X_{\setminus p}^{\text{ipl}}, \theta_p) f(\theta_p)$.

Schematic representation of the MCMC procedure

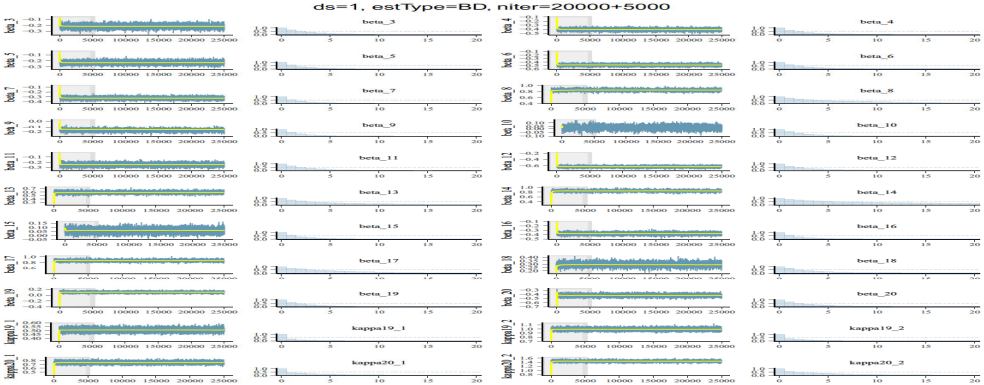
Initialize: $\gamma,\ \sigma^2,\ \upsilon^2,\ \xi = (\tilde{\alpha}, \tilde{\beta}),\ \tilde{\kappa},\ \theta,\ \omega,\ X_{\text{ipl}}$



Estimating functions and convergence behaviour

- Estimating functions are then approximated by means of corresponding sampling functions based on samples generated by MCMC algorithms from $f(X_{\text{ipl}}, \theta | X_{\text{obs}})$.
- The derivation of an estimating function is based on a loss function appropriate to the analysis objective; the decisive factor here is whether the individual missing values are considered to be the subject of inference.
- Statements on convergence are made in the literature in the context of the properties of MCMC methods, e.g. visually using autocorrelation and trajectory plots.
- The length of the necessary MCMC sequence lengths depends on the model; in the literature, 20,000 to 50,000 iterations and 20% as the initialisation phase are considered quite sufficient for complex latent hierarchical factor structures and mixed distribution modelling.

Autocorrelation function and trace plots



CART algorithm

- Classification and Regression Trees (CART) allow for many data constellations and variable types as a first modelling of the corresponding fully conditional densities, especially for discrete variables and cross-sectional data sets.
- Within the modelling of univariate fully conditional densities,
  - the specification of the set of conditional variables (inclusion of any transformations of variables, e.g. cross effects or higher powers) and
  - the specification of the hyperparameters.cross effects or higher powers), and
  - the specification of the hyperparameters.
- Recording of sufficient statistics of transformed variables.
- Drawing from the end nodes on the basis of the corresponding empirical distribution function using the Bayesian bootstrap.

CART algorithm (contd.)

- For metric variables, CART may need to be combined with kernel density estimates in order to partially remove the restriction of the approximation of the empirical distribution function to observed values within a subgroup.
- The consideration of hierarchical structures increases the computational effort; sufficient statistics may be available as a substitute.
- In principle, any complex modelling can be used to approximate univariate fully conditional density functions.
- CART itself can be characterised as a method of unsupervised learning.
- Supervision is carried out by specifying the set of conditional variables and the threshold values for model quality.

Modelling of restrictions

- Restriktionen $\sum_{r=1}^{R} X_{ir} \begin{matrix} \leq \\ \equiv \\ \geq \end{matrix} S_{iR}$ using data augmentation / slack variables

$$\tilde{S}_{iR} = \begin{cases} S_{iR} - \sum_{r=1}^{R+1} X_{ir} & \text{falls} \quad \geq, \\ 0 & \text{falls} \quad =, \\ -S_{iR} + \sum_{r=1}^{R+1} X_{ir} & \text{falls} \quad \leq \end{cases}$$

  and resulting carrier constraints.
- A priori, each of the variables or subset of variables occurring in the restriction can be the cause of the restriction violation.
- Generation of $R+1$ variable values and calculation of the remaining values.
  - The $R+1$ moves are discarded if the calculated values are not permissible.
  - Variation over the R possibilities to determine the set of $R+1$ variables.
- The report also includes the presentation of a joint modelling based on a degenerate multivariate normal distribution.

Modelling extensions

- Model extensions currently discussed in the literature, e.g. by combining mixed distribution modelling with spline regressions, can be introduced and used component by component.
- The presented framework using sequentially univariate density functions allows, in principle, the use of arbitrarily complex (univariate) modelling.
  - Consideration in the context of regression modelling, if target analysis can be delimited.
  - Use of factor analytical modelling as scalable models.
- The determination of the model or prediction quality can often not be done by means of statistics or quality criteria that can be calculated directly.
- Handling the trade-off between in-sample-fit and *computational burden* requires consistent benchmarking.

Specifics of variable types and data sets

- In principle, the modelling used should correspond to the scale level of the variables.
- If metric variables are present in the longitudinal data type, particular attention must be paid to the property of stationarity in the modelling used.
- Suitable stabilising transformations can be the use of growth rates or ratios.
- Each individual variable can be described by corresponding metadata (scale level – binary, ordinal, categorical, metric, categorisation of the number of values – countable, countably infinite, not countably infinite, stationarity – stationary, integration order, stabilising transformation, carrier – restricted, unrestricted).
- If necessary, sparse state-space representations are more expedient as a starting point for modelling than more complex modelling.

Model and prediction quality

- The starting point should be the consideration of a data scenario that is initially as complete as possible.
- The data set-specific establishment of quality thresholds in the context of pseudo-out-of-sample experiments based on cross-validation designs,
  - variable-specific minimum requirements for the prediction quality, respectively
  - variable-specific minimum requirements for model quality.
- The sequential approach allows for step-by-step expansion and further data types and modelling.
- Model complexity and flexibility are not a criterion in themselves.

Simulation study

- The basic setup consists of $J = 10$ variables for $N = 10000$ observations, with 3 restrictions (A, B, C), 2 of which are interlaced, i.e.

$$(1, 2, 3, (4), 5, 6), (7, 8, 9, 10)$$

				6					
			4			5		10	
1		2		3			7	8	9

$(A : V_{i1} + V_{i2} + V_{i3} = V_{i4})$, $(B : V_{i4} + V_{i5} = V_{i6})$, $(C : V_{i7} + V_{i8} + V_{i9} = V_{i10})$

- Simulation repetitions $S = 1000$ are considered, i.e. 1000 data sets are generated on the basis of the same data-generating process using the same population parameters.

Data generation

- Simulation of the data from 7-dimensional multivariate normal distribution, i.e.

$$V_r = (V_{i1}, V_{i2}, V_{i3}, V_{i5}, V_{i7}, V_{i8}, V_{i9})' \sim \mathcal{MVN}(\mu_r, \Sigma_r)$$

  mit $\mu_r = (100, 100, 100, 200, 100, 100, 100)'$ und $\Sigma_r = 20(.5\text{diag}(\iota_7) + .5\iota_7\iota_7')$ (ι_7 denotes $7\times$ 1 vector of ones).

- Operationalisation of the restrictions by calculating the variables V_4, V_6 and V_{10} as linear combinations (L) of the other variables, i.e.

$$V = LV_r \sim \mathcal{MVN}(\mu = L\mu_r, \Sigma = L\Sigma_r L')$$

  and the corresponding moments used in the benchmarking, i.e. μ, Σ and $q_x = \mu + \text{diag}(\Sigma)^{0.5}\Phi^{-1}$.i.e. μ, Σ and $q_x = \mu + \text{diag}(\Sigma)^{0.5}\Phi^{-1}(x)$, where $\Phi^{-1}(\cdot)$ represents the inverse cdf of the standard normal distribution.

Illustration of the data set structure

Five types of restriction violations, $\approx 5\%$ of the affected variables are missing for each type

lfd. Nr.	V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8	V_9	V_{10}
1	x	x	x	x	x	x	x	x	x	x
2	x	x	x	x	x	x	m	m	m	m
3	x	x	x	x	x	x	x	x	x	x
4	m	m	m	m	m	m	x	x	x	x
5	x	x	x	x	x	x	x	x	x	x
6	m	m	m	x	x	x	x	x	x	x
7	x	x	x	x	x	x	x	x	x	x
8	x	x	x	x	m	m	x	x	x	x
9	x	x	x	x	x	x	x	x	x	x
10	m	m	m	m	m	m	m	m	m	m
11	x	x	x	x	x	x	x	x	x	x
⋮										
N	x	x	x	x	x	x	x	x	x	x
# $\approx$ complete cases	85%	85%	85%	90%	85%	85%	90%	90%	90%	90%
					75%					

Handling of missing or implausible values

- Specification of the corresponding fully conditional densities taking into account the „identification problems“arising from the equality constraints, i.e.
  - $f(V_{i7}, \ldots, V_{i9}(, V_{i10})|V_{i1}, \ldots, V_{i3}(, V_{i4}), V_{i5}(, V_{i6}))$,
  - $f(V_{i1}, \ldots, V_{i3}(, V_{i4}), V_{i5}(, V_{i6})|V_{i7}, \ldots, V_{i9}(, V_{i10}))$,
  - $f(V_{i1}, V_{i2}(, V_{i3})|V_{i4}, V_{i5}(, V_{i6}), V_{i7}, V_{i8}, V_{i9}(, V_{i10}))$,
  - $f(V_{i5}(, V_{i6})|V_{i1}, \ldots, V_{i3}(, V_{i4}), V_{i7}, \ldots, V_{i9}(, V_{i10}))$,
  - $f(V_{i1}, V_{i2}, V_{i3}, (, V_{i4}), V_{i5}(, V_{i6}), V_{i7}, V_{i8}, V_{i9}(, V_{i10}))$.

Benchmarking

- relativer MSE als Gütekriterium, observed cases Schätzfunktionen ($\tilde{\theta}_s$) als Benchmark (naive + FIML? estimator),

$$\text{rel. MSE} = (1-)\frac{\frac{1}{S}\sum_{s=1}^{S}(\theta_s^* - \theta_{(s)})^2}{\frac{1}{S}\sum_{s=1}^{S}(\tilde{\theta}_s - \theta_{(s)})^2},$$

  where $\theta_s^* = \frac{1}{G}\sum_{g=1}^{G}\theta_{s,g}^*$ $(G = 10000/1000)$ and $\theta_{(s)}$ are referring to
  - the parameter value of the data-generating process (no variation over s), or
  - the estimation function based on the complete data (before deletion).
- parameter value of the data-generating process is either explicitly available (as functions of the parameters of the data-generating parametric density functions) or can be approximated as $\theta = \frac{1}{S}\sum_{s=1}^{S}\theta_s$, where θ_s represents the considered estimation function based on the complete data set (before deletion).

Operationalisation of the conditional densities

- multivariate conditional normal distributions (MVN), alternatively univariate normal distributions, i.e.

$$f(x_{i,\text{mis}}|x_{i,\text{obs}}) \propto |\Sigma_{\text{mis}|\text{obs}}|^{-.5} \exp\left\{-\frac{1}{2}(x_{i,\text{mis}} - \mu_{\text{mis}|\text{obs}})'\Sigma^{-1}_{\text{mis}|\text{obs}}(x_{i,\text{mis}} - \mu_{\text{mis}|\text{obs}})\right\},$$

where

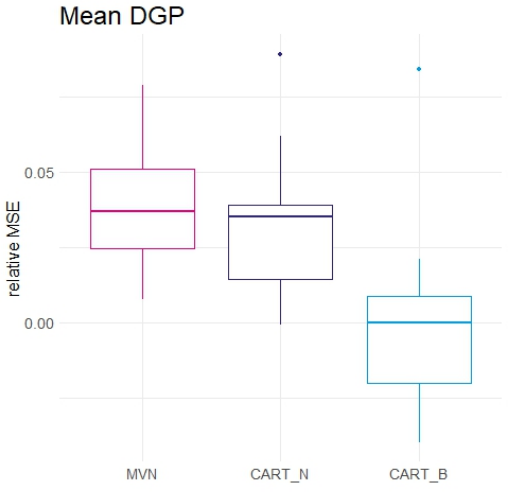
$$\mu_{\text{mis}|\text{obs}} = \mu_{\text{mis}} + \Sigma_{p,\text{mis,obs}}\Sigma^{-1}_{\text{obs}}(x_{i,\text{obs}} - \mu_{\text{obs}}) \quad \text{und} \quad \Sigma_{\text{mis}|\text{obs}} = \Sigma_{\text{mis}} - \Sigma_{p,\text{mis,obs}}\Sigma^{-1}_{\text{obs}}\Sigma'_{p,\text{mis,obs}}$$

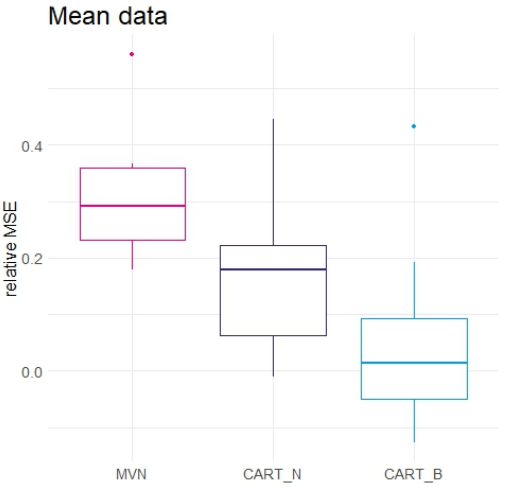
with

$$\mu = (\mu_{\text{mis}}, \mu_{\text{obs}})' \quad \text{und} \quad \Sigma = \begin{pmatrix} \Sigma_{\text{mis}} & \Sigma_{p,\text{mis,obs}} \\ \Sigma'_{p,\text{mis,obs}} & \Sigma_{\text{obs}} \end{pmatrix}.$$

- approximate using classification and regression trees (CART, at least 5 observations in end nodes, 10 observations for split, default values for further optimisation criteria)
  - Bayesian bootstrap based on the end node elements (non-parametric – CART_B),
  - normal distribution based on the end node elements (parametric – CART_N).

Simulation results – expected values





Simulation results – expected values

	DGP			DGPt			data		
	MVN	CART_N	CART_B	MVN	CART_N	CART_B	MVN	CART_N	CART_B
Min.	0,0078	-0,0008	-0,0398	0,0077	-0,0001	-0,0398	0,1797	-0,0103	-0,1268
1st Qu.	0,0246	0,0144	-0,0200	0,0246	0,0142	-0,0201	0,2322	0,0623	-0,0500
Median	0,0367	0,0350	0,0001	0,0367	0,0350	-0,0003	0,2910	0,1791	0,0143
Mean	0,0391	0,0342	0,0023	0,0390	0,0341	0,0022	0,3089	0,1672	0,0527
3rd Qu.	0,0508	0,0391	0,0089	0,0508	0,0389	0,0088	0,3586	0,2214	0,0926
Max.	0,0786	0,0890	0,0840	0,0786	0,0888	0,0835	0,5604	0,4461	0,4330

Note: Illustration of the relative MSE (1-rel. MSE) compared to benchmark estimators; distribution of parameter-specific values across the group of parameters considered.

Simulation results – minimum

	DGP			DGPt			data		
	MVN	CART_N	CART_B	MVN	CART_N	CART_B	MVN	CART_N	CART_B
Min.	0,1256	-0,4465	-0,4529	0,1256	-0,4465	-0,4529	0,0887	-4,8852	-4,3840
1st Qu.	0,1349	0,1000	0,0000	0,1349	0,1000	0,0000	0,1039	-0,2001	-0,0026
Median	0,1423	0,1735	0,0000	0,1423	0,1735	0,0000	0,1307	-0,0740	0,0000
Mean	0,1637	0,1063	-0,0406	0,1637	0,1063	-0,0406	0,1603	-0,5681	-0,4182
3rd Qu.	0,1986	0,2071	0,0064	0,1986	0,2071	0,0064	0,1721	0,0403	0,0000
Max.	0,2235	0,2324	0,0261	0,2235	0,2324	0,0261	0,3762	0,2243	0,2221

Note: Illustration of the relative MSE (1-rel. MSE) compared to benchmark estimators; distribution of parameter-specific values across the group of parameters considered.

Simulation results – maximum

	DGP			DGPt			data		
	MVN	CART_N	CART_B	MVN	CART_N	CART_B	MVN	CART_N	CART_B
Min.	0,1337	-0,4458	-0,8819	0,1337	-0,4458	-0,8819	0,0784	-5,9917	-8,0131
1st Qu.	0,1354	0,0961	0,0000	0,1354	0,0961	0,0000	0,1243	-0,2220	-0,0042
Median	0,1592	0,1692	0,0000	0,1592	0,1692	0,0000	0,1766	-0,1807	0,0000
Mean	0,1671	0,1032	-0,0843	0,1671	0,1032	-0,0843	0,1766	-0,6827	-0,7872
3rd Qu.	0,1961	0,2094	0,0043	0,1961	0,2094	0,0043	0,1959	0,0317	0,0000
Max.	0,2190	0,2311	0,0216	0,2190	0,2311	0,0216	0,3678	0,3243	0,1574

Note: Illustration of the relative MSE (1-rel. MSE) compared to benchmark estimators; distribution of parameter-specific values across the group of parameters considered.

Simulation results – variance/covariance

	DGP			DGPt			data		
	MVN	CART_N	CART_B	MVN	CART_N	CART_B	MVN	CART_N	CART_B
Min.	0,1042	-1,6477	-0,9647	0,1044	-1,6323	-1,0887	0,5311	-6,6447	-3,9709
1st Qu.	0,1500	0,0457	0,0109	0,1505	0,0617	-0,0295	0,5964	0,1749	-0,0023
Median	0,1625	0,1036	0,0924	0,1625	0,1094	0,0867	0,6520	0,4454	0,3355
Mean	0,1608	0,0146	0,0282	0,1608	0,0242	0,0058	0,6395	0,0640	0,0364
3rd Qu.	0,1756	0,1342	0,1404	0,1746	0,1369	0,1382	0,6892	0,5378	0,4553
Max.	0,1987	0,1762	0,1877	0,1987	0,1783	0,1816	0,7278	0,7104	0,6792

Note: Illustration of the relative MSE (1-rel. MSE) compared to benchmark estimators; distribution of parameter-specific values across the group of parameters considered.

Simulation results – quantiles

	DGP			DGPt			data		
	MVN	CART_N	CART_B	MVN	CART_N	CART_B	MVN	CART_N	CART_B
Min.	0,0681	-0,5086	-0,3859	0,0679	-0,4835	-0,3979	0,1157	-3,6348	-2,5414
1st Qu.	0,1011	0,0572	-0,0423	0,1011	0,0548	-0,0429	0,1716	-0,1085	-0,3704
Median	0,1329	0,0721	-0,0258	0,1328	0,0745	-0,0282	0,2053	-0,0212	-0,2885
Mean	0,1367	0,0397	-0,0452	0,1368	0,0405	-0,0478	0,2278	-0,2448	-0,4246
3rd Qu.	0,1583	0,0912	-0,0092	0,1581	0,0869	-0,0167	0,2497	0,0485	-0,1963
Max.	0,2492	0,1421	0,0521	0,2489	0,1306	0,0521	0,4849	0,3347	0,2039

Note: Illustration of the relative MSE (1-rel. MSE) compared to benchmark estimators; distribution of parameter-specific values across the group of parameters considered.

Simulation results - contd.

- Calculation times in R (Version 4.3.3, desktop computer with 64 GB RAM and 3.70 GHz, distributed over 8 cores)
  - MVN: 44.78 min
  - CART_N: 1.13 h
  - CART_B: 59.39 min
- The Gibbs principle has not yet been fully implemented, as no „update“of the parameter estimates was performed → Further improvements are possible.
- The generation of an a posteriori sample allows access to analytically not easily representable estimators, e.g. minimum.
- The generation of an a posteriori sample also allows „post-processing“in the sense of a reparametrisation in order to access optimal estimators taking into account the restrictions.
- Combination of CART and parametric drawing methods to approximate the fully conditional density depending on the data constellation makes sense.

- Replacing implausible values using the Bayesian paradigm, i.e. taking model uncertainty into account.
- Establishing automated handling of implausible values using delimitable data situations, e.g. cross-sectional data sets with many discrete variables.
- Extensions can then be made step by step in the context of a Bayesian approach using MCMC algorithms:
  - Extension to other data constellations.
  - Inclusion of more complex univariate fully conditional (hybrid) density functions.
- Monitoring of the structural properties of the occurrence of implausible values in new data sets.

Thank you for your attention!

Literature is listed within the references of the accompanying paper.

LIfBi
LEIBNIZ-INSTITUT FÜR BILDUNGSVERLÄUFE