Stable Graphical Model Estimation with Random Forests for Discrete, Continuous, and Mixed Variables

Bernd Fellinghauer\textsuperscript{a,b,*}, Peter Bühlmann\textsuperscript{b}, Martin Ryffel\textsuperscript{b}, Michael von Rhein\textsuperscript{c}, Jan D. Reinhardt\textsuperscript{a,d,e}

\textsuperscript{a}Swiss Paraplegic Research, Nottwil, Switzerland
\textsuperscript{b}Seminar for Statistics, ETH Zürich, Switzerland
\textsuperscript{c}Child Development Center, University Children’s Hospital, Zurich, Switzerland
\textsuperscript{d}Department of Health Sciences and Health Policy, University of Lucerne, Lucerne, Switzerland
\textsuperscript{e}International Society of Physical and Rehabilitation Medicine (ISPRM), Assenede (Ghent), Belgium

Abstract

A conditional independence graph is a concise representation of pairwise conditional independence among many variables. We propose Graphical Random Forests (GRaFo) for estimating pairwise conditional independence relationships among mixed-type, i.e. continuous and discrete, variables. The number of edges is a tuning parameter in any graphical model estimator and there is no obvious number that constitutes a good choice. Stability Selection helps choosing this parameter with respect to a bound on the expected number of false positives (error control).

We evaluate and compare the performance of GRaFo with Stable LASSO (StabLASSO), a LASSO-based alternative, across 5 simulated settings with \(p = 50, 100,\) and \(200\) variables, and we apply GRaFo to data from the Swiss Health Survey in order to evaluate how well we can reproduce the interconnection of functional health components, personal, and environmental factors, as hypothesized by the World Health Organization’s International Classification of Functioning, Disability and Health (ICF).

GRaFo performs well with mixed data and thanks to Stability Selection it provides an error control mechanism for false positive selection.

*Corresponding author. Swiss Paraplegic Research, Guido A Zäch-Str. 4, 6207 Nottwil, Switzerland. Tel. +41 41 939 6574; fax: +41 41 939 6577.
Email address: bernd.fellinghauer@paranet.ch (Bernd Fellinghauer)
1. Introduction

In many problems one is not confined to one response and a set of pre-
defined predictors. In turn, the interest is often in the association structure
of a whole set of $p$ variables, i.e. asking whether two variables are indepen-
dent conditional on the remaining $p-2$ variables. A conditional independence
graph (CIG) is a concise representation of such pairwise conditional indepen-
dence among many possibly mixed, i.e. continuous and discrete, variables. In
CIGs, variables appear as nodes, whereas the presence (absence) of an edge
among two nodes represents their dependence (independence) conditional on
all other variables. Applications include among many others also the study
of functional health (Strobl et al. (2009); Kalisch et al. (2010); Reinhardt
et al. (2011)).

We largely focus on the high-dimensional case where the number of vari-
ables (nodes in the graph) $p$ may be larger than sample size $n$. A popu-
lar approach to graphical modeling is based on the Least Absolute Shrink-
age and Selection Operator (LASSO, Tibshirani (1996)): see Meinshausen
and Bühlmann (2006) or Friedman et al. (2008) for the Gaussian case and
Ravikumar et al. (2010) for the binary case. However, empirical data of-
ten involve both discrete and continuous variables. Conditional Gaussian
distributions were suggested to model such mixed-type data with maximum
likelihood inference (Lauritzen and Wermuth (1989)), but no corresponding
high-dimensional method has been suggested yet. Dichotomization, though
always applicable, comes at the cost of lost information (MacCallum et al.
2002).

Tree-based methods are easy to use and accurate for dealing with mixed-
type data (Breiman et al. (1984)). Random Forests (Breiman (2001)) eval-
uate an ensemble of trees often resulting in notably improved performance
compared to a single tree (see also Amit and Geman (1997)). Furthermore,
permutation importance in Random Forests allows to rank the relevance of
predictors for one specific response. Since the definition of permutation im-
portance differs for discrete and continuous responses, ranking permutation
importances across responses of mixed-type is less obvious. However, such
ranking is essential to derive a network of the most relevant dependencies.
Stability Selection proposed by Meinshausen and Bühlmann (2010) is one possible framework to rank the edges in the CIG across different types of variables. In addition, it allows to specify an upper bound on the expected number of false positives, i.e. the falsely selected edges, and thus provides a means of error control.

We combine Random Forests estimation with appropriate ranking among mixed-type variables and error control from Stability Selection. We refer to the new method as Graphical Random Forests (GRaFo). The specific aims of the paper are a) to evaluate and compare the performance of GRaFo with Stable LASSO (StabLASSO), a LASSO-based alternative, across 5 simulated settings comprising different distributions for \( p = 50, 100, \) and 200 possibly mixed-type variables while sample size is \( n = 100, \) and b) to apply GRaFo to data from the Swiss Health Survey (SHS) to evaluate the interconnection of functional health components, personal, and environmental factors, as hypothesized by the World Health Organization’s (WHO) International Classification of Functioning, Disability and Health (ICF).

2. Graphical Modeling Based on Regression-Type Methods

2.1. Conditional Independence Graphs

Let \( X = \{X_1, \ldots, X_p\} \) be a set of (possibly) mixed-type random variables. The associated conditional independence graph of \( X \) is the undirected graph \( G_{\text{CIG}} = (\mathcal{V}, \mathcal{E}(G_{\text{CIG}})) \), where the nodes in \( \mathcal{V} \) correspond to the \( p \) variables in \( X \). The edges represent the pairwise Markov property, i.e. \( i \not\perp \!\!\!\!\perp j \notin \mathcal{E}(G_{\text{CIG}}) \) if and only if \( X_j \perp \!\!\!\!\perp X_i | X \setminus \{X_j, X_i\} \). For a rigorous introduction to graphical models, see, for example, the monographs by Whittaker (1990) or Lauritzen (1996).

We will now show that the pairwise Markov property can, under certain conditions, be inferred from conditional mean estimation.

**Theorem 1.** If \( X_j \) is a continuous random variable, assume that the conditional distribution of \( X_j \) given \( X \setminus \{X_j\} \) has a conditional density with respect to Lebesgue measure, and assume that \( \mathbb{E}[X_j|X \setminus \{X_j\}] \) exists for all \( j \). Furthermore, assume that the conditional density (for continuous random variable) or the conditional point probability (for discrete random variable) of \( X_j|X \setminus \{X_j\} \) is of the form

\[
f(x_j|X \setminus \{X_j\}) = f(x_j|\mathbb{E}[X_j|X \setminus \{X_j\}]) \quad \forall X \setminus \{X_j\}, \forall x_j \quad (C1)
\]
for some invertible function $h$. Then $X_j \perp X_i | X \setminus \{X_j, X_i\}$ if and only if $E[X_j | X \setminus \{X_j\}]$ is a function not depending on $X_j$ (i.e. $E[X_j | X \setminus \{X_j\}] = E[X_j | X \setminus \{X_j, X_i\}]$).

A proof is given in Section 7. Assumption (C1) trivially holds for a binary random variable $X_j$: for $x_j = 1$

$$f(x_j = 1 | X \setminus \{X_j\}) = P(X_j = 1 | X \setminus \{X_j\}) = E[X_j | X \setminus \{X_j\}]$$

and hence the function $h$ is the identity. Analogously, for a multinomial random variable $X_j$ with $C$ categories: for $x_j = (x_j^{(1)}, \ldots, x_j^{(C)})$ and $r = 1, \ldots, C$:

$$f_r(x_j | X \setminus \{X_j\}) = \pi_j^{(r)}(X \setminus \{X_j\})^{x_j^{(r)}} \{1 - \pi_j^{(r)}(X \setminus \{X_j\})\}^{1-x_j^{(r)}}$$

with $\pi_j^{(r)}(X \setminus \{X_j\}) = P[X_j^{(r)} = 1 | X \setminus \{X_j\}] = E[X_j^{(r)} | X \setminus \{X_j\}]$. Hence, (C1) holds with the identity function $h$. Moreover, if $(X_1, \ldots, X_p) \sim \mathcal{N}_p(\mathbf{0}, \Sigma)$, then (C1) holds as well (see for example [Lauritzen 1996]). However, for the conditional Gaussian case, we need to require for (C1) that the variance is fixed and is not depending on the variables we condition on. For example, let $X_1 \sim \mathcal{B}(1, \pi)$ be Bernoulli distributed and let

$$X_2 | X_1 \sim \begin{cases} \mathcal{N}(\mu_1, \sigma_1^2), & \text{if } X_1 = 1 \\ \mathcal{N}(\mu_2, \sigma_2^2), & \text{if } X_1 = 0 \end{cases}, \text{ where } \sigma_1^2 \neq \sigma_2^2.$$  

Then the distribution of $X_2 | X_1$ is not a function of the conditional mean alone.

Theorem 1 motivates our approach to infer conditional dependences, or edges in the CIG, via variable selection for many nonlinear regressions, i.e. determining whether a variable $X_i$ is relevant in $E[X_j | X \setminus \{X_j\}]$ (regression of $X_j$ versus all other variables).

### 2.2. Ranking Edges

In order to determine which edges should be included in the graphical model, the edges suggested by the individual regressions need to be ranked such that a smaller rank indicates a better candidate for inclusion. Depending on the assumed distribution of the responses and predictors, different measures are available. For instance, in ordinary least squares regression, where all variables are continuous, the size of the standardized regression coefficients is an obvious global ranking criterion. Analogously, when all variables...
are binary, coefficients from linear logistic regression lead to a global ranking. Note that each edge \(i - j\) is associated with two coefficients (\(X_j\) regressed on \(X_i\) and all other variables and vice versa for \(X_i\) on \(X_j\)). To be conservative, we rank each edge \(i - j\) relative to the smaller one of the two (absolute-valued) ranking coefficients.

If variables are mixed-type, a global ranking criterion might be impossible to find: Gaussian and non-Gaussian variables are not directly comparable. Instead, local rankings for each regression are performed separately ("local" means that we can rank the importance of predictors for every individual regression). Analogous to global ranking, each edge \(i - j\) is associated with two possible ranks and the worse among them is used.

When using Random Forests for performing the individual nonlinear regressions, the ranking scheme is obtained from Random Forests’ variable importance measure. When using the LASSO for individual linear or logistic regressions, the ranking scheme is obtained from absolute values of estimated regression coefficients.

We then have to decide on the number of edges to select, i.e. the tuning parameter. Assume it is given as \(q_{\text{thr}}\). Then for both global and local rankings we select the \(q_{\text{thr}}\) best-ranked edges across all \(p\) individual regressions. If this is impossible due to tied ranks, we neglect these tied edges and select only the remainder of edges not in violation of the threshold.

We next outline how Stability Selection can be used to guide the choice of \(q_{\text{thr}}\).

### 2.3. Aggregating Edge Ranks with Stability Selection

Stability Selection ([Meinshausen and Bühlmann (2010)](#)) allows the specification of an upper bound on the expected number \(\mathbb{E}[V]\) of false positives. It is based on subsampling ([Politis et al. (1999); Bühlmann and Yu (2002)](#)) random subsets \(X^{(1)}, \ldots, X^{(n_{\text{sub}})}\) of the original sample \(X_1, \ldots, X_n\), where each \(X^{(k)}\) contains \([n/2]\) sample points. Let \(\mathcal{E}(\hat{G}_{\text{CIG}}(X^{(k)}))\) denote the edges from a thresholded ranking based on \(X^{(k)}\), \(k = 1, \ldots, n_{\text{sub}}\). Stability Selection suggests to construct \(\mathcal{E}(\hat{G}_{\text{CIG}}(X))\), the set of all edges in the estimated CIG of \(X\), from all edges that were “sufficiently stable” across the \(n_{\text{sub}}\) subsets. More concretely, we choose only edges \(i - j\) which fulfill

\[
\frac{1}{n_{\text{sub}}} \sum_{k=1}^{n_{\text{sub}}} I\{i-j \in \mathcal{E}(\hat{G}_{\text{CIG}}(X^{(k)}))\} \geq \pi_{\text{thr}},
\]

(1)
where \( \pi_{thr} \in \left( \frac{1}{2}, 1 \right) \) imposes a threshold on the minimum relative frequency of edges across the \( n_{sub} \) subsets to be included in \( \mathcal{E}(\hat{G}_{\text{CIG}}(X)) \).

In their Theorem 1, Meinshausen and Bühlmann (2010) relate \( \mathbb{E}[V] \) to the maximum number of selected edges \( q_{thr} \) per subset (i.e. \( q_{thr} \) is the tuning parameter for thresholding the ranked edges), the number of possible edges \( p \cdot (p-1)/2 \) in \( \mathcal{E}(\hat{G}_{\text{CIG}}(X)) \), and the threshold \( \pi_{thr} \) from equation (1):

\[
\mathbb{E}[V] \leq \frac{q_{thr}^2}{(2\pi_{thr} - 1) \cdot p \cdot (p-1)/2}.
\]  

Both \( \mathbb{E}[V] \) and \( \pi_{thr} \) need to be specified a priori in order to determine \( q_{thr} \). As Meinshausen and Bühlmann (2010) argue, the choice of \( \pi_{thr} \) is of minor importance for a given \( \mathbb{E}[V] \), since a larger \( \pi_{thr} \) will mediate a larger \( q_{thr} \), and vice versa. We can thus use equation (2) to derive

\[
q_{thr} = \sqrt{(2\pi_{thr} - 1) \cdot \mathbb{E}[V] \cdot p \cdot (p-1)/2}.
\]

Henceforth, we fix \( n_{sub} = 100 \) and \( \pi_{thr} = 0.75 \), and vary \( \mathbb{E}[V] \) as required.

Note that equation (2) is based on two assumptions: 1) the estimation procedure is better than random guessing and 2) the probability of a false edge to be selected is exchangeable, i.e. each false edge is equally likely to be selected; for details we refer to Meinshausen and Bühlmann (2010).

3. Random Forests and LASSO Regression

3.1. Random Forests

Random Forests have, to date, not been used to estimate CIGs. They perform a series of recursive binary partitions of the data and construct the predictions from terminal nodes. Based on classification and regression trees (Breiman et al. (1984)) they allow convenient inference for mixed-type variables, also in the presence of interaction effects. Incorporating bootstrap (Efron (1979); Breiman (1996)) and random feature selection (Amit and Goldman (1997)), random subsets of both the observations and the predictors are considered. The relevance of each predictor can be assessed with permutation importance (Breiman (2002)), a measure of the error difference between a regular Random Forests fit and a Random Forests fit within which one predictor has been permuted at random to purge its relationship with the response. An implementation of Random Forests in R (R Development Core Team (2011)) is available in the randomForest package (Liaw and Wiener...
Since the goodness-of-fit of continuous and categorical responses is based on mean squared errors and majority votes, respectively, the goodness-of-fit and importance measures are not directly comparable across mixed-type responses. Thus a local ranking is derived, where each edge \( i - j \) is assigned either the rank of the permutation importance of predictor \( X^{(k)}_i \) for response \( X^{(k)}_j \) or of predictor \( X^{(k)}_j \) for response \( X^{(k)}_i \) (whichever is more conservative, i.e. assigns a worse rank) and finally aggregated with Stability Selection; the upper index \( (k) \) denotes the \( k^{th} \) subsample in Stability Selection. We refer to this procedure as Graphical Random Forests (GRaFo) henceforth.

### 3.2. Least Absolute Shrinkage and Selection Operator (LASSO)

In the case of linear regression for continuous responses and predictors, the LASSO (Tibshirani (1996)) penalizes with the \( \ell_1 \)-norm and corresponding penalty parameter \( \lambda \) the coefficients of some less relevant predictors to zero. The larger \( \lambda \) is chosen, the more coefficients will be set to zero. This concept has also been extended to logistic regression (Lokhorst (1999)) and implemented in R in the glmnet package (Friedman et al. (2010)). In the case of multinomial and mixed-type data, no eligible off-the-shelf implementations of the LASSO were available. We hence dichotomize these data according to a median split for continuous variables and aggregate categories such that the resulting frequency of the -1 and 1 categories was as balanced as possible for discrete variables.

CIG estimation via the LASSO with Stability Selection was suggested for Gaussian data by Meinshausen and Bühlmann (2010) and can be represented as a global ranking. For each response \( X^{(k)}_j \), we estimate LASSO regressions with all remaining \( X^{(k)} \setminus \{X^{(k)}_j\} \) as predictors and with a decreasing sequence of penalties \( \lambda^{(k),\max}, \ldots, \lambda^{(k),\min} \). Let \( \lambda^{(k)}_{ij} \) denote the largest penalty value of the sequence for which the coefficient of predictor \( X^{(k)}_i \) for response \( X^{(k)}_j \) is non-zero, and if no such penalty exists let \( \lambda^{(k)}_{ij} = 0 \). For each edge \( i - j \) we select the more conservative penalty \( \lambda^{(k)}_{i-j} = \min \left( \lambda^{(k)}_{ij}, \lambda^{(k)}_{ji} \right) \) and rank \( i - j \) relative to the global rank from the absolute-valued estimated regression coefficient corresponding to \( \lambda^{(k)}_{i-j} \). As before, the upper index \( (k) \) denotes the \( k^{th} \) subsample from Stability Selection. We denote this procedure in combination with Stability Selection as Stable LASSO (StabLASSO).
4. Simulation Study

4.1. Simulating Data from Directed Acyclic Graphs

We use a directed acyclic graph (DAG, cf., Whittaker (1990)) to embed conditional dependence statements among nodes representing the $p$ random variables. The associated CIG follows by moralization, i.e. connecting any two parents with a common child that are not already connected and removing all arrowheads (Lauritzen and Spiegelhalter (1988)).

Let $A$ be a $(p \times p)$-dimensional weight matrix with entries $a_{ij} \in \{[-1, -0.1] \cup \{0\} \cup [0.1, 1]\}$ if $i < j$ and $a_{ij} = 0$ otherwise. In addition, we sample $A$ to be sparse, i.e. we expect only one percent of its entries to deviate from 0. The non-zeros in $A$ encode the directed edges in a DAG we simulate from similarly as in Kalisch and Bühlmann (2007); see also Table 1. Furthermore, for all $i, j \in \{1, \ldots, p\}$ with $a_{ij} \neq 0$ let $u_{ij}$ and $v_{ij}$ be vectors that we use to impose some additional structure on multinomial variables: 1) at least one category of a multinomial predictor $X_i$ should have an effect opposite to the remainder, 2) the (total) effect of the categories of a multinomial predictor $X_i$ should be positive on some categories of a multinomial response $X_j$ and negative on others. For this purpose, we restrict $u_{ij} = (u_{ij}^{(1)}, \ldots, u_{ij}^{(C_i)})$ and $v_{ij} = (v_{ij}^{(1)}, \ldots, v_{ij}^{(C_j)})$:

$$u_{ij}^{(l)} \in \{-1, 1\} \; \forall \; l = 1, \ldots, C_i \; \text{s.t.} \; -C_i < \sum_{l=1}^{C_i} u_{ij}^{(l)} < C_i,$$

$$v_{ij}^{(s)} \in \{-1, 1\} \; \forall \; s = 1, \ldots, C_j \; \text{s.t.} \; -C_j < \sum_{s=1}^{C_j} v_{ij}^{(s)} < C_j.$$

With these definitions, we sample data from different distributions using the inverse link function to relate the conditional mean to all previously sampled predictors. Table 1 describes the settings in detail, covering models with purely Gaussian, purely Bernoulli, purely multinomial, and an alternating sequence of Gaussian and multinomial variables ("mixed" setting).

4.2. Simulating Data from the Ising Model

A common approach to model pairwise dependencies between a set of binary variables is the Ising model with probability function

$$p(x, \Theta) = \exp \left( \sum \theta_{ii} x_i + \sum \theta_{ij} x_i x_j - \Gamma(\Theta) \right)$$

(3)
| Distribution  | Model                                                                 | Conditional Mean                                                                 |
|--------------|----------------------------------------------------------------------|----------------------------------------------------------------------------------|
| Gaussian     | $X_j \sim \mathcal{N}(\mu_j, \sigma^2 = 1)$                        | $\mu_j = \sum_{i<j} a_{ij} x_i$                                                   |
| Bernoulli    | $X_j = 2\tilde{X}_j - 1,$                                           | $\mu_j = \sum_{i<j} a_{ij} x_i$                                                   |
|              | $\tilde{X}_j \sim \mathcal{B}(1, \pi_j)$                           | $\pi_j = \frac{\exp(\sum_{i<j} a_{ij} x_i)}{1+\exp(\sum_{i<j} a_{ij} x_i)}$ $\pi_j = \exp(\eta_j^{(s)}) / \sum_{r=1}^{C_j} \exp(\eta_j^{(r)})$ |
| Multinomial  | $X_j \sim \mathcal{M}(\pi_j = (\pi_j^{(1)}, \ldots, \pi_j^{(C_j)}))$| $\pi_j^{(s)} = \frac{\exp(\eta_j^{(s)})}{\sum_{r=1}^{C_j} \exp(\eta_j^{(r)})}$ |
|              | $\eta_j^{(s)} = \sum_{i<j} v_{ij}^{(s)} a_{ij} \sum_{l=1}^{C_i} u_{ij}^{(l)} (2I_{x_i=l} - 1),$ |                                                            |
|              | $C_j \sim \mathcal{U}\{3, 4, 5\},$                                | $\mu_j = \eta_j^{(1)}$, $\pi_j = \exp(\eta_j^{(s)}) / \sum_{r=1}^{C_j} \exp(\eta_j^{(r)})$ |
| Mixed        | $X_j \sim \begin{cases} \mathcal{N}(\mu_j, \sigma^2 = 1), & \text{if } j \notin \mathbb{N} \\ \mathcal{M}(\pi_j = (\pi_j^{(1)}, \ldots, \pi_j^{(C_j)})) & \text{else} \end{cases}$ | $\mu_j = \eta_j^{(1)}$, $\pi_j = \exp(\eta_j^{(s)}) / \sum_{r=1}^{C_j} \exp(\eta_j^{(r)})$ |
|              | $\eta_j^{(s)} = \sum_{i<j} v_{ij}^{(s)} a_{ij} x_i x_{ij} +$         |                                                            |
|              | $\sum_{i<i'j} v_{ij}'^{(s)} a_{ij} \sum_{l=1}^{C_i} u_{ij}^{(l)} (2I_{x_i=l} - 1)$ |                                                            |
|              | $C_j \sim \mathcal{U}\{3, 4, 5\},$                                | $\mu_j = \eta_j^{(1)}$, $\pi_j = \exp(\eta_j^{(s)}) / \sum_{r=1}^{C_j} \exp(\eta_j^{(r)})$ |

Table 1: The table shows the four simulation models based on DAGs. $\mathcal{N}$, $\mathcal{B}$, $\mathcal{M}$, and $\mathcal{U}$ are the Gaussian, Bernoulli, multinomial, and discrete uniform distribution, respectively. Initial values for $X_1$ are sampled with $\mu_1 = 0$, $\pi_1 = \frac{1}{2}$, and $\pi_1 = (\frac{1}{C_1}, \ldots, \frac{1}{C_1})$, respectively, where $C_1 \sim \mathcal{U}\{3, 4, 5\}$. The weights $a_{ij}$ are chosen from $\{-1, -0.1\} \cup \{0\} \cup [0.1, 1]$ to determine the dependence relationships among the random variables. The scalars $u_{ij}^{(l)}$ and $v_{ij}^{(s)}$ are chosen from $\{-1, 1\}$ to impose additional structures on multinomial random variables.
for realizations $\mathbf{x} \in \mathbf{X}$, normalization constant $\Gamma(\Theta)$, and $(p \times p)$-dimensional symmetric parameter matrix $\Theta = \{\theta_{ij}\}_{i,j\in\{1,\ldots,p\}}$. From the conditional densities of equation (3) it follows that $\theta_{ij} = 0 (\theta_{ij} \neq 0)$ implies the absence (presence) of edge $i - j$ in the associated CIG. See also Ravikumar et al. (2010).

We sample the diagonal and the upper-triangular matrix of $\Theta$ uniformly from $\{-1, 0, 1\}$ such that the average neighborhood size for each node equals 4. The lower-triangular matrix equals its upper counterpart. We use the Gibbs sampler (cf., Givens and Hoeting (2005)) to sample realizations from equation (3). Höfling and Tibshirani (2009) provide an implementation in the BMN package in R.

4.3. Simulation Results

For $p \in \{50, 100, 200\}$ variables and samples of size $n = 100$, each of the 5 simulation models was averaged over 50 repetitions. The results are shown in Figures [1-6]. Error control for small bounds on the expected number of false positives $\mathbb{E}[V]$ could be achieved for both GRaFo and StabLASSO in all but the mixed setting with $p = 200$ in Figure [6].

In the Gaussian, Bernoulli and Ising settings, StabLASSO seems to perform slightly better than GRaFo for small error bounds and rather similar across the figures for the true/false positive rates (third column of Figures [1-6]). Note that StabLASSO sets many coefficients to 0. As a consequence, a large proportion of edges cannot be selected for false positive rates smaller than 1 resulting in some StabLASSO curves not covering the entire range of the rates.

In the multinomial and mixed setting, GRaFo returned satisfactory results whilst dichotomization resulted in a dramatic loss in performance of StabLASSO (cf., MacCallum et al. (2002); Altman and Royston (2006); Royston et al. (2006)). In general, both procedures seem to perform best in the Gaussian setting, followed by the mixed, multinomial, Bernoulli, and Ising setting, respectively. The latter seems especially hard for both procedures if the upper error bound in equation (2) for $\mathbb{E}[V]$ is chosen small. Nevertheless, given one’s willingness to expect more errors, the rate figures indicate the potential to recover (parts of) the true structure (cf., Ravikumar et al. (2010); Höfling and Tibshirani (2009)). The “raw” counterparts, Random Forests and LASSO, seem to perform quite similar to GRaFo and StabLASSO across all settings, but they do not provide any error measure or guidance how to regularize the procedure.
A violation of condition (C1) of Theorem 1 in the mixed setting could explain the failure of both GRaFo and StabLASSO to achieve error control for \( p = 200 \). However, both the mixed setting with \( p = 50 \) and \( p = 100 \) returned very few observed errors and remained well below the error bounds indicating the problematic behavior may be linked to larger values of \( p \). Also, for any setting it is unlikely that the exchangeability assumption holds. [Meinshausen and Bühlmann (2010)] argue that Stability Selection appears to be robust to violations, but did not study mixed data which may be particularly affected.

The computational cost is growing rather quickly with growing \( p \). The runtime of a single of the 50 repetitions per setting is in the order of 15 minutes for GRaFo and 20 minutes for StabLASSO for \( p = 50 \) and increases to several hours for GRaFo and 30 minutes for StabLASSO in the case of \( p = 200 \). Each batch of 50 repetitions was run in parallel on 50 cores of the BRUTUS high-performance cluster comprising quad-core AMD Opteron 8380 2.5 Ghz CPUs with 1 GB of RAM per core using the Rmpi package [Yu (2010)] available in R.

5. Functional Health in the Swiss General Population

5.1. The Importance of Functional Health

According to the World Health Organization’s (WHO) new framework of the International Classification of Functioning, Disability and Health (ICF, [WHO (2001)]) the lived experience of health ([Stucki et al. (2008)]) can be structured in experiences related to body functions and structures as well as to activity and participation in society. All of these are, in turn, influenced by a variety of so-called personal factors such as gender, income, or age and environmental factors including individual social relations and supports as well as properties of larger macro social systems such as the economy (see Figure [7]). Also, the WHO and The World Bank recommend in their recent World Report on Disability (2011) that functional health state descriptors are analyzed in conjunction with other health outcomes and, particularly, that more research is conducted on “[...] the interactions among environmental factors, health conditions, and disability [...]” ([WHO and The World Bank (2011) p. 267]). Under these prerequisites it is of interest which variables are conditionally dependent on each other. For instance, “Does the income distribution affect participation, conditional on known impairments, environmental, and personal factors?”.
Figure 1: The rows correspond to the Gaussian, Bernoulli, and Ising model with \( p = 50 \). Their true CIGs have 16, 16 and 89 edges, respectively. The first two columns report the observed number of true and false positives ("o") relative to the bound in eqn. [2] for the expected number \( E[V] \) of false positives ("\( \frac{1}{2} \)"") for GRaFo and StabLASSO, respectively, averaged over 50 simulations. The third column reports the averaged true and false positive rates of GRaFo and StabLASSO relative to the performance of their “raw” counterparts without Stability Selection.
Figure 2: The rows correspond to the Gaussian, Bernoulli, and Ising model with $p = 100$. Their true CIGs have 58, 58 and 182 edges, respectively. The first two columns report the observed number of true and false positives (“o”) relative to the bound in eqn. [2] for the expected number $E[V]$ of false positives (“”) for GRaFo and StabLASSO, respectively, averaged over 50 simulations. The third column reports the averaged true and false positive rates of GRaFo and StabLASSO relative to the performance of their “raw” counterparts without Stability Selection.
Figure 3: The rows correspond to the Gaussian, Bernoulli, and Ising model with $p = 200$. Their true CIGs have 334, 334 and 369 edges, respectively. The first two columns report the observed number of true and false positives (“o”) relative to the bound in eqn. (2) for the expected number $E[V]$ of false positives (“i”) for GRaFo and StabLASSO, respectively, averaged over 50 simulations. The third column reports the averaged true and false positive rates of GRaFo and StabLASSO relative to the performance of their “raw” counterparts without Stability Selection.
Figure 4: The rows correspond to the multinomial and mixed-type model with $p = 50$. Their true CIGs both have 16 edges. The first two columns report the observed number of true and false positives (“o”) relative to the bound in eqn. (2) for the expected number $E[V]$ of false positives (“|”) for GRaFo and StabLASSO, respectively, averaged over 50 simulations. The third column reports the averaged true and false positive rates of GRaFo and StabLASSO relative to the performance of their “raw” counterparts without Stability Selection.
Figure 5: The rows correspond to the multinomial and mixed-type model with $p = 100$. Their true CIGs both have 58 edges. The first two columns report the observed number of true and false positives (“o”) relative to the bound in eqn. (2) for the expected number $\mathbb{E}[V]$ of false positives (“|”) for GRaFo and StabLASSO, respectively, averaged over 50 simulations. The third column reports the averaged true and false positive rates of GRaFo and StabLASSO relative to the performance of their “raw” counterparts without Stability Selection.
Figure 6: The rows correspond to the multinomial and mixed-type model with \( p = 200 \). Their true CIGs both have 334 edges. The first two columns report the observed number of true and false positives ("\( o \)") relative to the bound in eqn. (2) for the expected number \( \mathbb{E}[V] \) of false positives ("\( \]") for GRaFo and StabLASSO, respectively, averaged over 50 simulations. The third column reports the averaged true and false positive rates of GRaFo and StabLASSO relative to the performance of their “raw” counterparts without Stability Selection.
Figure 7: The International Classification of Functioning, Disability and Health (ICF) model relates aspects of human functioning and provides a common language for practitioners.
5.2. Study Population

We use GRaFo for a secondary analysis of cross-sectional observational
data on functional health from the Swiss Health Survey (SHS) in 2007. Data
were obtained from the Federal Statistics Office of Switzerland. The original
study was based on a stratified random sample of all private Swiss households
with fixed line telephones. Within each household one household member
aged 15 or older was randomly selected. The survey was completed by a
total of 18760 persons, corresponding to a participation rate of 66 percent
(Graf (2010)). The mean age of study participants was 49.6 years (±18.5).
The data were mostly collected with computer assisted telephone interviews.
Further information is available elsewhere (Storni (2011)).

5.3. Variables

The SHS included various information on symptoms (in particular pain),
impairments, and activity limitations. Since the respective items were some-
times nominal, sometimes ordinal, and sometimes (e.g. body mass index)
metric, we dichotomized each item so that 1 was indicative of having any
kind of problem. As overall summary scores on functioning and disability
were not recommendable (Reinhardt et al. (2010)), we followed the frame-
work of the WHO’s biopsychosocial model of health, outlined in the ICF
(WHO (2001), see Figure 7), and other theoretical considerations (WHO
and The World Bank (2011); Reinhardt et al. (2010)) in constructing sum
indices (see Table 2). The plausibility of all indices was checked using the
Stata 11 confirmatory factor analysis module confa (Kolenikov (2009)). In
each case the index construction was tested and the null hypothesis of a
diagonal structure of the covariance matrix rejected.

We created a dummy variable for labor market participation restrictions
such that 1 identified persons who gave up work, reduced the number of
working hours, or changed jobs because of health reasons. We also created a
dummy variable for participation in leisure physical activity (LPA) differen-
tiating between people participating in leisure activities leading to sweating
at least once a week and those who do not. General health perception was
measured with the following question and answer options: “How would you
rate your health in general? Very good, good, fair, poor, or very poor?”.
We further included indicators of socio-economic status (SES) in our anal-
ysis: equivalence household income, years of formal education, employment
status, and migration background (foreign origin of at least one parent). On
the macro- or cantonal-level we obtained information on the Swiss counties’
| Construct       | Variable specification                                                                 | Coding for ‘yes’ |
|-----------------|-----------------------------------------------------------------------------------------|-----------------|
| Impairment      | Problems with vision 1 (any problem)                                                    | 1               |
|                 | Problems with hearing 1 (any problem)                                                   | 1               |
|                 | Problems with speaking 1 (any problem)                                                  | 1               |
|                 | Body mass index over 30 or under 16 1                                                  | 1               |
|                 | Urinary incontinence 1 (any problem)                                                    | 1               |
|                 | Defecation problems 1 (any problem)                                                     | 1               |
|                 | Feeling weak, tired, lack of energy 1 (any problem)                                     | 1               |
|                 | Sleeping problems 1 (any problem)                                                       | 1               |
|                 | Tachycardia 1 (any problem)                                                             | 1               |
| Range of sum index: 0-9 |                                                                                       |                 |
| Pain            | Pain in head 1 (any pain)                                                               | 1               |
|                 | Pain in chest 1 (any pain)                                                              | 1               |
|                 | Pain in stomach 1 (any pain)                                                            | 1               |
|                 | Pain in back 1 (any pain)                                                               | 1               |
|                 | Pain in hands 1 (any pain)                                                              | 1               |
|                 | Pain in joints 1 (any pain)                                                             | 1               |
| Range of sum index: 0-6 |                                                                                       |                 |
| Activity limitation | Problems with walking 1 (any problem)                                                   | 1               |
| Social support  | Feeling lonely 0                                                                        | 0               |
|                 | Missing someone to turn to 0                                                            | 0               |
|                 | Having at least one supportive family member 1                                          | 1               |
|                 | Having someone to turn to 1                                                             | 1               |
| Range of sum index: 0-4 |                                                                                       |                 |
| Social network utilization | At least weekly visits of family 1                                                      | 1               |
|                 | At least weekly phone calls with family 1                                               | 1               |
|                 | At least weekly visits of friends 1                                                     | 1               |
|                 | At least weekly phone calls with friends 1                                              | 1               |
|                 | At least weekly participation in clubs/associations/parties 1                          | 1               |
| Range of sum index: 0-5 |                                                                                       |                 |

Table 2: Construction rules of sum indices for functioning (pain, impairment, activity limitation) and social integration (social support and social network utilization) from 37 dichotomous variables.

(cantons) gross domestic products (GDP), Gini coefficients, and crime rates for 2006. Moreover, we considered information on gender, age, marital status (being married), alcohol consumption (in grams per day), and current smoking (yes/no).

Of these, in total, 20 mixed-type variables (see Table 3), income had the highest number of missing values with roughly 6 percent. Overall, less than 0.85 percent of replies were missing corresponding to 2687 cases with one or more missing values. To assess their effect, we estimated the CIG once with casewise deletion and once with imputation of missing values with the `missForest` procedure ([Stekhoven and Bühlmann](2011)) available in R.
| Type                  | Variable                              | % Missing |
|-----------------------|---------------------------------------|-----------|
| > 2 categories        | Impairment index                      | 5.92      |
|                       | Pain index                            | 0.37      |
|                       | Activity limitation index             | 0.69      |
|                       | Social support index                  | 5.84      |
|                       | Social network utilization index       | 2.32      |
|                       | General health perception             | 0.05      |
| Dichotomous           | Male                                  | 0.00      |
|                       | Married                               | 0.09      |
|                       | Paid work                             | 0.03      |
|                       | Migration background                  | 4.73      |
|                       | Smoker                                | 0.07      |
|                       | Work restriction                      | 0.00      |
|                       | Leisure physical activity             | 0.00      |
| Continuous            | Age                                   | 0.00      |
|                       | Years of formal education             | 0.07      |
|                       | Income                                | 5.94      |
|                       | Alcohol consumption (in grams per day)| 2.59      |
|                       | Gross domestic product                | 0.00      |
|                       | Gini coefficient                      | 0.00      |
|                       | Crime rate                            | 0.00      |

Table 3: List of all 20 variables used in the CIG estimation, their type, and their percentage of missing values.
Figure 8: Conditional independence graph of the $p = 20$ variables (nodes) remaining after construction of indices based on the 2007 Swiss Health Survey estimated with GRaFo. Edges were selected with respect to an upper bound of 5 on the expected number of false positives, see equation (2). Five nodes (social network utilization, migration background, smoker, work restriction, and LPA) were isolated (no edges) and thus neglected.

5.4. Research Hypothesis

From the WHO’s ICF model [WHO (2001), see Figure 7], we hypothesized that all variables on functional and general health perception, and all variables on social status, networks, and supports were connected via paths within the same component of the CIG.

5.5. Findings

Figure 8 shows the resulting graph from our application of GRaFo to the (non-imputed) data on functional health from the SHS with casewise deletion of missing values regularized for a bound (as in equation (2)) for an expected number of false positives $E[V] \leq 5$. The selected edge sets for the imputed and casewise deleted data were quite similar for various bounds on $E[V]$ and even identical for $E[V] \leq 5$ (not shown). In the following, we thus focus on the CIG derived from the complete observations remaining after casewise deletion of missing values.
The resulting edges for $E[V] \leq 1$ depict relatively obvious associations known from everyday observations. Interestingly, general health perception is conditionally dependent on activity limitation but conditionally independent of impairment and pain. In the larger graph for $E[V] \leq 5$, one sees that general health perception, impairments, and pain are connected through a path of several environmental and personal factors such as social support, being married, age, etc. That implies, for instance, that we do not need information on impairment to predict general health perception if we have information on activity limitation and the remaining predictors, whereas activity limitation is an essential predictor of general health perception even if information on all the remaining predictors is provided. For instance, a person with a spinal cord injury who has no activity limitation because of social and technological supports, could thus still report good health. This finding is supported by other sources reporting that many people with disabilities do not consider themselves to be unhealthy [WHO and The World Bank (2011); Watson (2002)]. In the 2007-2008 Australian National Health Survey, 40 percent of people with a severe or profound impairment rated their health as good, very good, or excellent [Australian Bureau of Statistics (2009)].

As regards our hypothesis derived from the ICF model [WHO (2001)], we can confirm that the bulk of individual level variables form one component and support the biopsychosocial model of health: Functional and general health influence each other and are connected with a variety of environmental and personal factors. However, not all candidate personal and environmental factors were related in our study. This may be due to our conservative upper bound on the error that is likely to favor false negatives, i.e. missing edges. There may also be an issue with our selection of variables that was restricted by the choices of the original survey team. In particular, macro-level variables pertaining information about the counties, in which the individuals are nested, form a second component. It may be that their effect is already contained in the individual-level variables, for example paid work. Five variables do not appear in the graph entirely: social network utilization, migration background, smoker, work restriction, and LPA. If we remove the three macro-level variables GDP, Gini, and crime rate from the model, the connectivity of the individual-level component does not change. Instead, the two variables migration background and social network utilization are now present as a separate component (not shown).

Unfortunately, lack of information on the directions of relationships is a
weakness of CIGs. Also, condition (C1) of Theorem 1 and the exchangeability condition have likely been violated. Regardless, given the high face validity of the findings and the achievement of error control in the mixed setting for small $p$ in Section 4.3, the results seem satisfactory.

The runtime of GRaFo depends also on $n$, even if $p$ is small. Hence, estimation of the SHS graph was executed in parallel on 10 cores of the BRUTUS cluster with a runtime of roughly 8 hours.

6. Conclusion

We propose GRaFo (Graphical Random Forests) performed satisfactory, mostly on par or superior to StabLASSO, LASSO, and Random Forests. Error control could be achieved in all but the mixed-type simulation with $p = 200$. Violation of assumption (C1) in Theorem 1 and of the exchangeability condition might be responsible for this behavior. In contrast, in most of the other settings GRaFo was very conservative and observed errors were well below their expected upper bound. The Ising model, the sole model not based on DAGs, was particularly hard for both GRaFo and StabLASSO resulting in few true positives if error bounds were chosen very small.

Poor results for the LASSO in the multinomial and mixed case may be improved by feasible modifications of the LASSO, such as an extension of the group LASSO (Meier et al. (2008)) to multinomial responses. However, penalization among different types of variables (including the issue of scaling) is not a straightforward task.

The Swiss Health Survey graph consists of an individual- and a macro-level variable cluster which were highly stable with respect to the way of handling missing values. Exclusion of the macro-level cluster did not affect the individual-level cluster. For a small error bound, our hypothesis that all factors should connect could not be fully confirmed, though a strong tendency toward the ICF’s biopsychosocial model of health was evident in the individual-level cluster.

7. Proof of Theorem 1

We have to show that for the density or point probability $f$:

$$X_j \perp X_i | X \setminus \{X_j, X_i\} \iff \mathbb{E}[X_j | X \setminus \{X_j\}] = \mathbb{E}[X_j | X \setminus \{X_j, X_i\}].$$

(4)
We first show "⇒" of eqn. (4):

\[ X_j \perp \perp X_i \mid \mathbf{X} \setminus \{X_j, X_i\} \]
\[ \Rightarrow f(x_j | \mathbf{X} \setminus \{X_j\}) = f(x_j | \mathbf{X} \setminus \{X_j, X_i\}) \forall X \setminus \{X_j\}, \forall x_j \]
\[ \Rightarrow \int x_j f(x_j | \mathbf{X} \setminus \{X_j\}) dx_j = \int x_j f(x_j | \mathbf{X} \setminus \{X_j, X_i\}) dx_j \forall X \setminus \{X_j\} \]
\[ \Rightarrow \mathbb{E}[X_j | \mathbf{X} \setminus \{X_j\}] = \mathbb{E}[X_j | \mathbf{X} \setminus \{X_j, X_i\}] \forall X \setminus \{X_j\}. \]

We now show "⇐" of eqn. (4):

\[ \mathbb{E}[X_j | \mathbf{X} \setminus \{X_j\}] = \mathbb{E}[X_j | \mathbf{X} \setminus \{X_j, X_i\}] \forall X \setminus \{X_j\}. \]
\[ \Rightarrow (c1) \quad f(x_j | \mathbf{X} \setminus \{X_j\}) = f(x_j | \mathbf{X} \setminus \{X_j, X_i\}) \forall X \setminus \{X_j\}, \forall x_j \]
\[ \Rightarrow X_j \perp \perp X_i \mid \mathbf{X} \setminus \{X_j, X_i\}. \]

Thus equation (4) holds.

8. Acknowledgment

The authors would like to thank Gerold Stucki, Markus Kalisch, Philipp Rütimann, and Holger Höfling for valuable feedback and discussion.

9. References

References

Altman, D.G., Royston, P., 2006. The cost of dichotomising continuous variables. Brit Med J 332, 1080.

Amit, Y., Geman, D., 1997. Shape quantization and recognition with randomized trees. Neural Comput 9, 1545–1588.

Australian Bureau of Statistics, 2009. National Health Survey: Summary of Results, 2007-2008. Australian Bureau of Statistics, Canberra.

Breiman, L., 1996. Bagging predictors. Mach Learn 24, 123–140.

Breiman, L., 2001. Random Forests. Mach Learn 45, 5–32.

Breiman, L., 2002. Setting Up, Using, And Understanding Random Forests V4.0.
Breiman, L., Friedman, J., Olshen, R., Stone, C., 1984. Classification and Regression Trees. Wadsworth, Inc., California.

Bühlmann, P., Yu, B., 2002. Analyzing bagging. Ann Stat 30, 927–961.

Efron, B., 1979. Bootstrap methods: Another look at the jackknife. Ann Stat 7, 1–26.

Friedman, J., Hastie, T., Tibshirani, R., 2008. Sparse inverse covariance estimation with the graphical Lasso. Biostatistics 9, 432–441.

Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. J Stat Softw 33, 1–22.

Givens, G.H., Hoeting, J.A., 2005. Computational Statistics. Hohn Wiley & Sons, Inc., New Jersey.

Graf, E., 2010. Rapport de méthodes. Enquête suisse sur la santé 2007. Plan d’échantillonnage, pondérations et analyses pondérées des données. Office Fédéral de la Statistique, Neuchâtel.

Höfling, H., Tibshirani, R., 2009. Estimation of sparse binary pairwise markov networks using pseudo-likelihoods. J Mach Learn Res 10, 883–906.

Kalisch, M., Bühlmann, P., 2007. Estimating high-dimensional directed acyclic graphs with the PC-algorithm. J Mach Learn Res 8, 613–636.

Kalisch, M., Fellinghauer, B., Grill, E., Maathuis, M.H., Mansmann, U., Bühlmann, P., Stucki, G., 2010. Understanding human functioning using graphical models. BMC Med Res Methodol 10, 14.

Kolenikov, S., 2009. Confirmatory factor analysis using confa. Stata J 9, 329–373.

Lauritzen, S.L., 1996. Graphical Models. Oxford University Press, Oxford.

Lauritzen, S.L., Spiegelhalter, D.J., 1988. Local computations with probabilities on graphical structures and their application to expert systems (with discussion). J Roy Stat Soc B 50, 157–224.
Lauritzen, S.L., Wermuth, N., 1989. Graphical models for associations between variables, some of which are qualitative and some quantitative. Ann Stat 17, 31–57.

Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R News 2, 18–22.

Lokhorst, J., 1999. The Lasso and generalised linear models. Honors project. The University of Adelaide, Australia.

MacCallum, R.C., Zhang, S., Preacher, K.J., Rucker, D.D., 2002. On the practice of dichotomization of quantitative variables. Psychol Methods 7, 19–40.

Meier, L., van de Geer, S., Bühlmann, P., 2008. The group Lasso for logistic regression. J Roy Stat Soc B 70, 53–71.

Meinshausen, N., Bühlmann, P., 2006. High-dimensional graphs and variable selection with the Lasso. Ann Stat 34, 1436–1462.

Meinshausen, N., Bühlmann, P., 2010. Stability selection (with discussion). J Roy Stat Soc B 72, 417–473.

Politis, D.N., Romano, J.P., Wolf, M., 1999. Subsampling. Springer, Berlin.

R Development Core Team, 2011. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. ISBN 3-900051-07-0.

Ravikumar, P., Wainwright, M.J., Lafferty, J.D., 2010. High-dimensional Ising model selection using $\ell_1$-regularized logistic regression. Ann Stat 38, 1287–1319.

Reinhardt, J.D., Fellinghauer, B., Strobl, R., Stucki, G., 2010. Dimension reduction in human functioning and disability outcomes research: Graphical models versus principal components analysis. Disabil Rehabil 32, 1000–1010.

Reinhardt, J.D., Mansmann, U., Fellinghauer, B., Strobl, R., Grill, E., von Elm, E., Stucki, G., 2011. Functioning and disability in people living with spinal cord injury in high- and low-resourced countries: A comparative analysis of 14 countries. Int J Public Health 56, 341–352.
Royston, P., Altman, D.G., Sauerbrei, W., 2006. Dichotomizing continuous predictors in multiple regression: A bad idea. Stat Med 25, 127–141.

Stekhoven, D.J., Bühlmann, P., 2011. MissForest - nonparametric missing value imputation for mixed-type data. Preprint. arXiv: 1106.2068v1.

Storni, M., 2011. Enquêtes, sources: Enquête suisse sur la santé. Office Fédéral de la Statistique, Neuchâtel.

Strobl, R., Stucki, G., Grill, E., Müller, M., Mansmann, U., 2009. Graphical models illustrated complex associations between variables describing human functioning. J Clin Epidemiol 62, 922–933.

Stucki, G., Kostanjsek, N., Üstün, B., Cieza, A., 2008. ICF-based classification and measurement of functioning. Eur J Phys Rehabil Med 44, 315–328.

Tibshirani, R., 1996. Regression shrinkage and selection via the Lasso. J Roy Stat Soc B 58, 267–288.

Watson, N., 2002. Well, I know this is going to sound very strange to you, but I don’t see myself as a disabled person: Identity and disability. Disability & Society 17, 509–527.

Whittaker, J., 1990. Graphical Models in Applied Multivariate Statistics. John Wiley & Sons, Inc., New Jersey.

WHO, 2001. International Classification of Functioning, Disability and Health (ICF). WHO Press, Geneva.

WHO and The World Bank, 2011. World Report on Disability. WHO Press, Geneva.

Yu, H., 2010. Rmpi: Interface (Wrapper) to MPI (Message-Passing Interface).