Bayesian Analysis (2011) **6**, Number 3, pp. 429–456

Goodness-of-Fit of Conditional Regression Models for Multiple Imputation

Stefano Cabras*, María Eugenia Castellanos† and Alicia Quirós‡

Abstract. We propose the calibrated posterior predictive p-value (*cppp*) as an interpretable goodness-of-fit (GOF) measure for regression models in sequential regression multiple imputation (SRMI). The *cppp* is uniformly distributed under the assumed model, while the posterior predictive p-value (*ppp*) is not in general and in particular when the percentage of missing data, *pm*, increases. Uniformity of *cppp* allows the analyst to evaluate properly the evidence against the assumed model. We show the advantages of *cppp* over *ppp* in terms of power in detecting common departures from the assumed model and, more importantly, in terms of robustness with respect to *pm*. In the imputation phase, which provides a complete database for general statistical analyses, default and improper priors are usually needed, whereas the *cppp* requires a proper prior on regression parameters. We avoid this problem by introducing the use of a minimum training sample that turns the improper prior into a proper distribution. The dependency on the training sample is naturally accounted for by changing the training sample at each step of the SRMI. Our results come from theoretical considerations together with simulation studies and an application to a real data set of anthropometric measures.

Keywords: Calibrated posterior predictive p-value, Discrepancy measure, Minimum training sample, Missing at random, Predictive distribution, Sequential regression multiple imputation

1 Introduction

Multiple imputation (MI) techniques, first introduced by Rubin (1978), have become popular in the last decades, and nowadays there are a variety of multiple imputation models and software available (e.g., the `MICE` package in `R`). MI consists of filling missing data values of a variable with multiple samples from an imputation model. In practice, MI is a simulation technique in which the missing values are replaced by $S > 1$ draws from the conditional predictive distribution of the imputed variable given the others in the data set. Imputation techniques include, among others, ordinary least-squares regression, logistic regression, factor analysis, variance components estimation, and proportional hazard models. Further details on MI can be found in Rubin (2004, 1996); Schafer (1999) and references therein.

*Dipartimento di Matematica e Informatica, Universitá Degli Studi di Cagliari, Italy, mailto:s.cabras@unica.it

†Departamento de Estadística e I.O., Universidad Rey Juan Carlos, Spain, mailto:maria.castellanos@urjc.es

‡Departamento de Estadística e I.O., Universidad Rey Juan Carlos, Spain, mailto:alicia.quiros@urjc.es

 DOI:10.1214/11-BA617

In this paper we assume that missing values are generated under the missing at random (MAR) mechanism, in which the probability of missingness depends only on available information. Based on this, we focus on linear regression models for imputation. In particular, we consider sequential regression multiple imputation (SRMI) that imputes each variable in turn, depending on the rest, using a regression model (Raghunathan et al. 2001). It is important to note that in SRMI, once a variable is completed, it is used as a regressor for the next variable to be imputed, in a Gibbs-like manner, until convergence of the regression coefficients is achieved. As SRMI operates at the level of the conditional distribution of a variable given the rest, then the joint distribution for all variables is not specified.

Despite the popularity of MI methods, assessing their goodness-of-fit (GOF) is not a common practice and only a few papers address this problem. Gelman et al. (2005), Gelman (2004) and Abayomi et al. (2008) propose Bayesian posterior predictive checks for imputed data sets. Graphical diagnostics and exploratory data analysis are considered in Gelman et al. (2005) and Gelman (2004) whereas Abayomi et al. (2008) judge the propriety of the imputed values by comparison with the observed data, using Kolmogorov-Smirnov (KS) tests for each variable, together with bivariate scatterplots and residual plots.

In He et al. (2007), the authors suggest the use of the posterior predictive p-value, *ppp*, originally proposed by Rubin (1984) and Meng (1994), further formalized and extended in Gelman et al. (1996), to assess the GOF of a certain imputation parametric model. Let Y be the variable to be imputed following the sampling model $f(y|\boldsymbol{\beta})$ where $\boldsymbol{\beta} \in \Theta$ is distributed according to prior $\pi(\boldsymbol{\beta})$, then the *ppp* is defined as

$$ppp(\boldsymbol{y}) = \Pr\left(D(\boldsymbol{Y}^{\text{rep}}, \boldsymbol{\beta}) \geq D(\boldsymbol{y}, \boldsymbol{\beta})|\boldsymbol{y}\right). \tag{1}$$

Here $D(\boldsymbol{y}, \boldsymbol{\beta})$ is a discrepancy measure, $\boldsymbol{y}$ represents the observed data and the distribution of $\boldsymbol{Y}^{\text{rep}}$ is the posterior predictive distribution, $m(\boldsymbol{Y}|\boldsymbol{y}) = \int_{\Theta} f(\boldsymbol{y}|\boldsymbol{\beta})\pi(\boldsymbol{\beta}|\boldsymbol{y})d\boldsymbol{\beta}$. Without loss of generality, in (1) we assume that larger values of D indicate incompatibility. The *ppp* is usually approximated by a Monte Carlo sum where $\boldsymbol{\beta}$ and $\boldsymbol{Y}^{\text{rep}}$ are drawn from the posterior distribution $\pi(\boldsymbol{\beta}|\boldsymbol{y})$ and $f(\boldsymbol{Y}|\boldsymbol{\beta})$, respectively.

In this work, we focus on a GOF approach based on p-values, pointing out that *ppp* cannot be interpreted under the usual uniform distribution in (0,1), as also noted in Bayarri and Berger (2000); Dahl (2006); Hjort et al. (2006). We show here that *ppp* is conservative for GOF of SRMI, when GOF is based on suitable discrepancy measures. The work of Robins et al. (2000) demonstrated that *ppp* is asymptotically conservative when using GOF statistics whose distribution depends on unknown parameters. In order to overcome these drawbacks we propose to assess the GOF of SRMI with the calibrated posterior predictive p-value, *cppp*, proposed in Hjort et al. (2006). In that work, the authors propose post-processing the *ppp* obtaining the *cppp*, defined as

$$cppp(\boldsymbol{y}) = \Pr\left(ppp(\boldsymbol{Y}) \leq ppp(\boldsymbol{y})\right), \tag{2}$$

where $\boldsymbol{Y}$ comes from the prior predictive distribution, $m(\boldsymbol{Y}) = \int_{\Theta} f(\boldsymbol{y}|\boldsymbol{\beta})\pi(\boldsymbol{\beta})d\boldsymbol{\beta}$. In contrast to *ppp*, *cppp* is uniformly distributed under the null model, but it requires proper

priors for $\boldsymbol{\beta}$, being sensitive to the choice of $\pi(\boldsymbol{\beta})$. Nonetheless, Bayesian imputation techniques usually make use of default and improper priors. This creates a problem for GOF based on *cppp* that we overcome by introducing the use of a minimal training sample. Following Definition 1 in Berger and Pericchi (1996) we make use of the idea of minimum training sample, that consists of a sample of the data, of minimal size, that trains an improper prior into a proper distribution through the Bayes' theorem. Additionally, we compare this approach with that based on conjugate priors.

Finally, as stated in Hjort et al. (2006), the performance of the GOF critically depends on the choice of $D(\boldsymbol{y}, \boldsymbol{\beta})$. We choose several types of discrepancy measures, $D(\boldsymbol{y}, \boldsymbol{\beta})$, intuitively related to regression models, but a more comprehensive study of such discrepancies is beyond the scope of this paper. Also note that this paper is focused on GOF rather than on model selection and we encourage the reader to look at O'Hagan (2003) for a discussion of the role of Bayesian model checking versus model selection.

The rest of the paper is organized as follows: GOF for SRMI, using two different priors and several discrepancy measures, is explained in Section 2. Section 3 validates the performance of the proposed technique with a simulation study and Section 4 illustrates an application to a real data set. Further remarks and conclusions are contained in Section 5.

2 Goodness-of-fit for SRMI

Consider the incomplete dataset, $(\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_Q)$, where variables are ordered by increasing number of missing values. SRMI imputes each variable, one at a time, given the rest in a sequence of S imputations. Let $s = 1, \ldots, S$ denote the step of the procedure, the imputation model for $\boldsymbol{Y}_q$, $q = 1, \ldots, Q$, is the posterior predictive distribution, $m(\boldsymbol{Y}_q|\boldsymbol{y}_q)$, based on the regression model

$$\boldsymbol{Y}_q|\boldsymbol{Y}_{-q}, \boldsymbol{\beta} \sim N_n(\boldsymbol{Y}_{-q}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}), \quad \boldsymbol{\beta} = (\boldsymbol{\beta}, \sigma^2) \in \mathbb{R}^{p_q} \times \mathbb{R}^+, q = 1, \ldots, Q, \tag{3}$$

where $\boldsymbol{Y}_{-q}$ denotes the rest of the variables or a subset of these, and the intercept, while N_n represents the n-dimensional normal distribution. For the sake of simplicity, regression parameters are denoted by $\boldsymbol{\beta}$ instead of $\boldsymbol{\beta}_q$. Notice that $\boldsymbol{Y}_{-q}$ includes a total of $p_q \leq Q$ variables that are either fully observed or have been completed in previous steps. In order to evaluate the GOF of the conditional regression models in SRMI, we propose the *cppp* as a measure of the adequacy of each regression model (3) used to impute variable $\boldsymbol{Y}_q$ at each step s. In the sequel, we illustrate the details of the procedure.

As stated in the previous section, *cppp* requires a proper prior distribution on $\boldsymbol{\beta}$. In order to evaluate the robustness of *cppp* with respect to $\pi(\boldsymbol{\beta})$, we consider two different choices: *i*) a trained prior, $\pi^t(\boldsymbol{\beta})$, that is based on the usual default prior trained with a minimum training sample (Berger and Pericchi 1996), drawn from the observed data; and *ii*) a vague conjugate prior, $\pi^c(\boldsymbol{\beta})$, in which the analyst specifies the order of vagueness.

2.1 Trained prior and posterior distribution

Suppose we are imputing, at a certain step s, variable $\boldsymbol{Y} = \boldsymbol{Y}_q$ using model (3) where $\boldsymbol{Y}_{-q}$ is here denoted by $\boldsymbol{X}$. Here $\boldsymbol{Y}$ is a vector of length n and $\boldsymbol{X}$ is an $n \times p$ matrix with $p = p_q$. We assume $\boldsymbol{Y}$ and $\boldsymbol{X}$ partitioned as $\boldsymbol{Y} = (\boldsymbol{y}_o^T, \boldsymbol{Y}_m^T)^T$ and $\boldsymbol{X} = (\boldsymbol{X}_o^T, \boldsymbol{X}_m^T)^T$, where subindices o and m indicate the n_o observed and $n_m = n - n_o$ missing values of $\boldsymbol{Y}$. Note that $\boldsymbol{X}$ is either fully observed or previously completed in the sequential imputation scheme.

Let

$$\pi^N(\boldsymbol{\beta}) = \pi^N(\boldsymbol{\beta}, \sigma^2) \propto \frac{1}{\sigma^2}$$

be the usual default prior for parameters $\boldsymbol{\beta}$ in the linear regression model. The trained prior is

$$\pi^t(\boldsymbol{\beta}) \propto f(\boldsymbol{y}_t|\boldsymbol{X}_t, \boldsymbol{\beta})\pi^N(\boldsymbol{\beta})$$

where $f(\boldsymbol{y}_t|\boldsymbol{X}_t, \boldsymbol{\beta})$ is the density of model (3) and $\boldsymbol{y}_t$ is a random sample of size n_t drawn from the observed data $\boldsymbol{y}_o$, and $\boldsymbol{X}_t$ represents the corresponding rows of $\boldsymbol{X}$. In the case of the linear regression model, with p covariates, the minimum training sample $\{\boldsymbol{y}_t, \boldsymbol{X}_t\}$ that turns $\pi^N(\boldsymbol{\beta})$ into a proper density has size $n_t = p + 1$.

The trained prior, $\pi^t(\boldsymbol{\beta})$, is given by

$$\begin{aligned} \boldsymbol{\beta} \,|\, \boldsymbol{y}_t, \boldsymbol{X}_t, \sigma^2 &\sim N_p(\widehat{\boldsymbol{\beta}}_t, \sigma^2 V_\beta^{-1}) \\ \sigma^2 \,|\, \boldsymbol{y}_t, \boldsymbol{X}_t &\sim \text{Inv}-\chi^2(n_t - p, \widehat{\sigma}_t^2) = \text{Inv}-\chi^2(1, \widehat{\sigma}_t^2) \end{aligned} \tag{4}$$

with

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_t &= (\boldsymbol{X}_t^T \boldsymbol{X}_t)^{-1} \boldsymbol{X}_t^T \boldsymbol{y}_t \\ V_\beta &= (\boldsymbol{X}_t^T \boldsymbol{X}_t) \\ \widehat{\sigma}_t^2 &= (\boldsymbol{y}_t - \boldsymbol{X}_t \widehat{\boldsymbol{\beta}_t})^T (\boldsymbol{y}_t - \boldsymbol{X}_t \widehat{\boldsymbol{\beta}_t}). \end{aligned}$$

Let $\boldsymbol{y}_{o\setminus t} = \boldsymbol{y}_o \setminus \boldsymbol{y}_t$ and $\boldsymbol{X}_{o\setminus t} = \boldsymbol{X}_o \setminus \boldsymbol{X}_t$ be the observed data after removing $\{\boldsymbol{y}_t, \boldsymbol{X}_t\}$, then the posterior distribution, $\pi(\boldsymbol{\beta}|\boldsymbol{y}_o, \boldsymbol{X}_o)$, is the result of the usual conjugate analysis with response $\boldsymbol{y}_{o\setminus t}$ and design matrix $\boldsymbol{X}_{o\setminus t}$ and the trained prior distribution $\pi^t(\boldsymbol{\beta})$ (4). Notice that the trained posterior, $\pi(\boldsymbol{\beta}|\boldsymbol{y}_o, \boldsymbol{X}_o)$, equals the posterior distribution calculated over all observed data with default prior $\pi^N(\boldsymbol{\beta})$. An advantage of $\pi^t(\boldsymbol{\beta})$ is that it allows us to use a default improper prior $\pi^N(\boldsymbol{\beta})$ in the GOF. Nevertheless, the main criticism of trained priors is that $\pi^t(\boldsymbol{\beta})$ depends on the training sample. It is important to stress here that this dependency is naturally accounted for by changing $\{\boldsymbol{y}_t, \boldsymbol{X}_t\}$ in each imputation step of the SRMI.

2.2 Conjugate prior and posterior distribution

The conjugate prior, $\pi^c(\boldsymbol{\beta})$, for (3) is

$$\begin{aligned} \boldsymbol{\beta} \,|\, \sigma^2 &\sim N_p(\boldsymbol{\beta}_0, \sigma^2 V_0^{-1}) \\ \sigma^2 &\sim \text{Inv}-\chi^2(n_0, \sigma_0^2) \end{aligned} \tag{5}$$

and the posterior distribution, $\pi^c(\boldsymbol{\beta}|\boldsymbol{y}_o, \boldsymbol{X}_o)$, is the usual one for the conjugate analysis, where prior parameters in (5) must be specified reflecting the analyst's uncertainty about $\boldsymbol{\beta}$. Here N_p denotes a p-dimensional normal distribution.

Note that, different to the case of a non-informative prior, both the posterior $\pi^c(\boldsymbol{\beta}|\boldsymbol{y}_o, \boldsymbol{X}_o)$ and the GOF, based on *cppp*, depend on the information contained in $\pi^c(\boldsymbol{\beta})$. In Section 3 the effect of the degree of vagueness of the prior in the performance of the *cppp* is analyzed and both priors, $\pi^c(\boldsymbol{\beta})$ and $\pi^t(\boldsymbol{\beta})$, are compared.

2.3 Discrepancy measures

In this section we specify the discrepancy measures used to compute (1) for each s. Although many discrepancy measures may be used to assess the GOF of an imputation model, we consider the following discrepancies which are intuitively related to regression and are based on residuals, $e_i = y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}$, for $i = 1, \ldots, n$:

$$R^2 = 1 - \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2}, \quad SSR = \sum_{i=1}^n \frac{e_i^2}{\sigma^2},$$

$$Max = \max_i \frac{|e_i|}{\sigma}, \qquad KS = \text{KS discrepancy for normality of } e_1, \ldots, e_n,$$

where y_i is the i-th element of $\boldsymbol{Y}$, $\boldsymbol{x}_i$ is the corresponding row of $\boldsymbol{X}$ and $\bar{y} = \sum_{i=1}^n y_i/n$. These discrepancies are random variables defined over $\boldsymbol{Y}$, i.e. $D((\boldsymbol{y}_o, \boldsymbol{Y}_m), \boldsymbol{\beta})$. As R^2 has a different interpretation with respect to the rest of measures, in the sense that lower values of R^2 indicate poorer fit, we use the following definition of *ppp*: $\Pr\left(R^2(\boldsymbol{Y}^{\text{rep}}, \boldsymbol{\beta}) \leq R^2(\boldsymbol{y}, \boldsymbol{\beta})|\boldsymbol{y}\right)$. Note that all these quantities are discrepancy measures and not statistics, implying, for instance, that R^2 may take negative values.

2.4 SRMI and approximation of *cppp*

Consider again $(\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_Q)$ ordered by increasing number of missing values. In the first round of SRMI, $s = 1$, the first variable with missing values, $\boldsymbol{Y}_r$, is imputed using its posterior predictive distribution $m(\boldsymbol{Y}_r|\boldsymbol{y}_{-r})$ based on model (3) where, in this case, the covariates $\boldsymbol{Y}_{-r} = (\boldsymbol{1}, \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{r-1})$ are the intercept and the fully observed variables, if any. Note that if $r = 1$, $\boldsymbol{Y}_{-1}$ consists only of the intercept term. The rest of the variables, $\boldsymbol{Y}_{r+1}, \ldots, \boldsymbol{Y}_Q$, are imputed sequentially using as covariates the fully observed variables jointly with those previously imputed. For $s > 1$, the imputation process is carried out as in round 1, except that, in each regression, all other variables are included as predictors. Schematically, for round $s > 1$, and for each q such that $r \leq q \leq Q$, impute variable $\boldsymbol{Y}_q$ by drawing missing values from $m(\boldsymbol{Y}_q|\boldsymbol{y}_{-q})$ for model (3), where

$$\boldsymbol{Y}_{-q} = \boldsymbol{Y}_{-q}^s = (\boldsymbol{1}, \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{r-1}, \boldsymbol{Y}_r^s \ldots, \boldsymbol{Y}_{q-1}^s, \boldsymbol{Y}_{q+1}^{s-1}, \ldots, \boldsymbol{Y}_Q^{s-1})$$

to obtain the completed qth variable at step s, namely $\boldsymbol{Y}_q^s$.

For fixed s and q in the above SRMI procedure, let $\boldsymbol{Y} = \boldsymbol{Y}_q$, $\boldsymbol{X} = \boldsymbol{Y}_{-q}$ and denote,

with an abuse of notation, $ppp = ppp^{s,q}$ and $cppp = cppp^{s,q}$. The following algorithm provides an approximation of *ppp* and *cppp* when using the trained prior.

1. Draw a minimum training sample, $\{\boldsymbol{y}_t, \boldsymbol{X}_t\}$, where $\boldsymbol{y}_t$ is a random sample of size $n_t = p + 1$ uniformly drawn from the observed data $\boldsymbol{y}_o$ (see Section 2.1), and $\boldsymbol{X}_t$ are the corresponding rows of $\boldsymbol{X}$.

2. Calculate $ppp(\boldsymbol{y}_o)$ using the following Monte Carlo sum:

$$ppp(\boldsymbol{y}_o) = \frac{1}{J}\sum_{j=1}^{J} I\{D(\boldsymbol{Y}_j^{\text{rep}}, \boldsymbol{\beta}_j) \geq D(\boldsymbol{y}_o, \boldsymbol{Y}_{m,j}, \boldsymbol{\beta}_j)\}, \tag{6}$$

   where $\boldsymbol{\beta}_j$ is drawn from $\pi(\boldsymbol{\beta}|\boldsymbol{y}_o, \boldsymbol{X}_o)$ defined in Section 2.1, while imputed values, $\boldsymbol{Y}_{m,j}$, and replicated data, $\boldsymbol{Y}_j^{\text{rep}}$, are simulated from model (3) given $\boldsymbol{\beta}_j$ with covariates $\boldsymbol{X}_m$ and $\boldsymbol{X}$, respectively.

3. Approximate *cppp* according to:

$$cppp(\boldsymbol{y}_o) = \frac{1}{K}\sum_{k=1}^{K} I\{ppp(\boldsymbol{Y}_{o,k}) \leq ppp(\boldsymbol{y}_o)\}, \tag{7}$$

   where $\boldsymbol{Y}_{o,k}$ is drawn from model (3) with covariates $\boldsymbol{X}_o$, and $\boldsymbol{\beta}_k$ is simulated from $\pi^t(\boldsymbol{\beta})$. For each $\boldsymbol{Y}_{o,k}$, compute $ppp(\boldsymbol{Y}_{o,k})$ using step 2.

In the case of using a conjugate prior, in order to approximate *ppp* and *cppp*, remove step 1 and replace $\pi^t(\boldsymbol{\beta})$ and $\pi(\boldsymbol{\beta}|\boldsymbol{y}_o, \boldsymbol{X}_o)$ with $\pi^c(\boldsymbol{\beta})$ and $\pi^c(\boldsymbol{\beta}|\boldsymbol{y}_o, \boldsymbol{X}_o)$, respectively.

We propose to assess the GOF of SRMI, for a certain variable $\boldsymbol{Y}_q$, using the whole sequence $cppp^{1,q}, \ldots, cppp^{S,q}$, as illustrated in the application. We suggest discarding the whole SRMI if there is at least one q such that the corresponding model is not compatible with the observed data. In fact, due to the nature of SRMI, the imputation of variable Y_q affects the imputation of the rest of variables.

It is worth remarking here that we are assessing the GOF of conditional regression models and it is theoretically possible that the imputation procedure may not converge to a stationary distribution, because the conditional densities may not be compatible with any joint distribution of all variables (Gelman and Speed 1993). This problem is beyond the scope of the paper and our GOF procedure is not able to detect any such kind of incompatibility even if all conditional models were compatible. However, as noted in Raghunathan et al. (2001), this rarely occurs in practical cases. Moreover, according to van Buuren (2007), the approach of Fully Conditional Specification, that includes SRMI, should be preferred to that of Joint Modeling when the joint distribution of the data cannot easily be specified.

3 Simulation study

In this section we present the results of a simulation study in which we investigate and compare the performance of *ppp* and *cppp* using $\pi^t(\boldsymbol{\beta})$ and some choices of $\pi^c(\boldsymbol{\beta})$.

Firstly, we consider one variable with missing values, $\boldsymbol{y}$, and one complete covariate, $\boldsymbol{x}$, and, secondly, we explore the case where both $\boldsymbol{y}$ and $\boldsymbol{x}$ are incomplete. We generate n observations according to three models:

Null model: $y_i|x_i \sim N(1 + x_i, 1)$;

Alternative 1: $y_i|x_i \sim N(1 + x_i^2, 1)$;

Alternative 2: $y_i|x_i \sim N(1 + z_i, 1)$, where $\boldsymbol{z}$ is another covariate such that $\boldsymbol{z} \neq \boldsymbol{x}$,

for $i = 1, \ldots, n$. Covariates $\boldsymbol{x}$ and $\boldsymbol{z}$ are generated independently from two standard normal distributions. In order to mimic a MAR mechanism, let pm denote the proportion of missing values, then we randomly delete $[n \cdot pm]$ elements of $\boldsymbol{y}$ and $\boldsymbol{x}$, with probability of missingness proportional to $\Phi(\boldsymbol{x})$ and $\Phi(\boldsymbol{y})$, respectively, where Φ denotes the cumulative standard normal distribution. The above MAR mechanism corresponds, under Alternative 2, to an MCAR mechanism for $\boldsymbol{y}$ and $\boldsymbol{x}$.

Within these frameworks we consider assessing the GOF of the following imputation model:

$$y_i|x_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2), \quad \text{for the case of missing observations only in } \boldsymbol{y}.$$

In the case of two incomplete variables, the two models derived from SRMI definition are:

$$\begin{aligned} y_i|x_i &\sim N(\beta_0 + \beta_1 x_i, \sigma^2) \\ x_i|y_i &\sim N(\beta_0' + \beta_1' y_i, \sigma'^2). \end{aligned}$$

This study has been performed with different sample sizes, $n \in \{100, 500\}$, and proportion of missing values, $pm \in \{0.1, 0.4, 0.6, 0.9\}$. For each scenario, we evaluate the null sampling distribution of *ppp* and *cppp*, using the four discrepancy measures defined above, and their behavior under alternatives 1 and 2, using 100 replications of data.

In the case of missing observations only in $\boldsymbol{y}$, we consider the improper prior $\pi^N(\boldsymbol{\beta}) = \pi^N(\beta_0, \beta_1, \sigma^2)$ joint with the training procedure, where the training sample, $\boldsymbol{y}_t$, changes for each *cppp*. We compare this with three different conjugate priors $\pi^c(\boldsymbol{\beta}) = \pi^c(\beta_0, \beta_1, \sigma^2)$ with $n_0 = 2$, 0.2 and 0.1, $\sigma_0^2 = 1$, $\boldsymbol{\beta}_0 = (0, 0)$ and $V_0^{-1} = 100\boldsymbol{I}$, in (5). Such parameters have been chosen in order to evaluate the effect of vagueness. In particular, note that the moments of $\pi^c(\boldsymbol{\beta})$ are not defined for $n_0 \leq 2$. When $\boldsymbol{x}$ and $\boldsymbol{y}$ are both incomplete, we only consider the trained prior.

Figure 1 shows the sampling distributions of *cppp* and *ppp* under the null model, with $n = 100$ and $pm = 0.1$ and 0.6, considering the trained prior when only $\boldsymbol{y}$ is incomplete. As expected, the *cppp* is uniformly distributed under the null model, while this generally does not apply to *ppp*. In particular, for discrepancies R^2 and SSR, *ppp* concentrates around 0.5 (plots in the left), while for KS and Max it loses uniformity when pm increases (plots in the right). The same comments apply to all considered versions of $\pi^c(\boldsymbol{\beta})$, showing that under the null model, the automatic procedure, based

on $\pi^N(\boldsymbol{\beta})$ provides similar results to that based on $\pi^c(\boldsymbol{\beta})$. By construction, sampling null distributions of the *cppp* are uniform, therefore they are so when both $\boldsymbol{x}$ and $\boldsymbol{y}$ are incomplete (not shown here).

Therefore, it is more interesting to compare the power of *ppp* and *cppp* when rejecting the null model for a cut-off of 0.05 under alternatives 1 and 2. Power, under each alternative, is approximated by the proportion of the 100 p-values below 0.05. Figure 2 represents the approximated power for alternatives 1 and 2 using $\pi^N(\boldsymbol{\beta})$ and $\pi^c(\boldsymbol{\beta})$ with $n_0 = 0.2$, for $n = 100$ and values of pm ranging between 0.1 and 0.9 and only missing values in $\boldsymbol{y}$. For Alternative 1, KS and Max are the most powerful discrepancies followed by R^2 and SSR, with the corresponding *cppp* showing better performance than *ppp*. Under Alternative 2, the most powerful discrepancy is R^2 and again, it is more powerful under *cppp* than under *ppp*. When using the vague conjugate prior, power of both measures increases slightly.

For both alternatives, power generally increases with n, see Figure 3. As expected, the larger the pm, the smaller the power, specially in the case of *ppp* whose robustness with respect to pm is considerably lower than that of *cppp*, in particular for reasonable pm, i.e. $pm < 0.9$. Due to the missing mechanism, for $pm = 0.9$, data loses its quadratic structure in Alternative 1, leading to a sensible loss of power. Instead, when assuming the MCAR mechanism in Alternative 1, data keep the original quadratic structure also for $pm = 0.9$ and the power of *cppp* rises.

Figures 4-5 report the power of *cppp* under the trained priors for the imputation models of $\boldsymbol{y}|\boldsymbol{x}$ and $\boldsymbol{x}|\boldsymbol{y}$, respectively. For Alternative 1, in spite of the larger overall number of missing values, when both $\boldsymbol{x}$ and $\boldsymbol{y}$ are missing, powers for the imputation model of $\boldsymbol{y}|\boldsymbol{x}$ are essentially in line with those in Figures 2-3. Regarding $\boldsymbol{x}|\boldsymbol{y}$, the lower power, compared to that of $\boldsymbol{y}|\boldsymbol{x}$, is due to the quadratic relationship we suppose between $\boldsymbol{y}|\boldsymbol{x}$. However, the power's decrease for $\boldsymbol{x}|\boldsymbol{y}$ is abundantly compensated with the power's increase for $\boldsymbol{y}|\boldsymbol{x}$ and thus the joint assessment of the GOF of the linear imputation model would strongly indicate its inadequacy. The increase of the power, at $pm = 0.4$ of KS and Max for $\boldsymbol{x}|\boldsymbol{y}$, is due to the combination of the quadratic relationship $\boldsymbol{y}|\boldsymbol{x}$ and the assumed MAR mechanism. In fact, after deleting observations, there remain some points, corresponding to large $\boldsymbol{y}$, that act as outliers inducing linear model rejection. For large pm such outliers disappear resulting in model compatibility. However, also in such situations, using *ppp*, instead of *cppp*, we cannot detect model incompatibility. For Alternative 2, as the overall number of missing values increases, power decreases more than when only $\boldsymbol{y}$ is missing, specially for $n = 500$.

The above results show that the power of *cppp* depends on the prior used and the global amount of missing values in the data. However, differences between priors seem to vanish for large sample sizes. The smaller power of the *ppp* using $\pi^N(\boldsymbol{\beta})$ is due to less weight of this prior with respect to data, when compared with $\pi^c(\boldsymbol{\beta})$. Using $\pi^c(\boldsymbol{\beta})$ instead of $\pi^N(\boldsymbol{\beta})$ we are measuring not only the discrepancy of the imputation model with data, but also between the assumed imputation model and the prior. The same reason can be ascribed to the behavior of *cppp*: when it is based on the trained prior its power decreases compared to the corresponding one based on the conjugate prior.

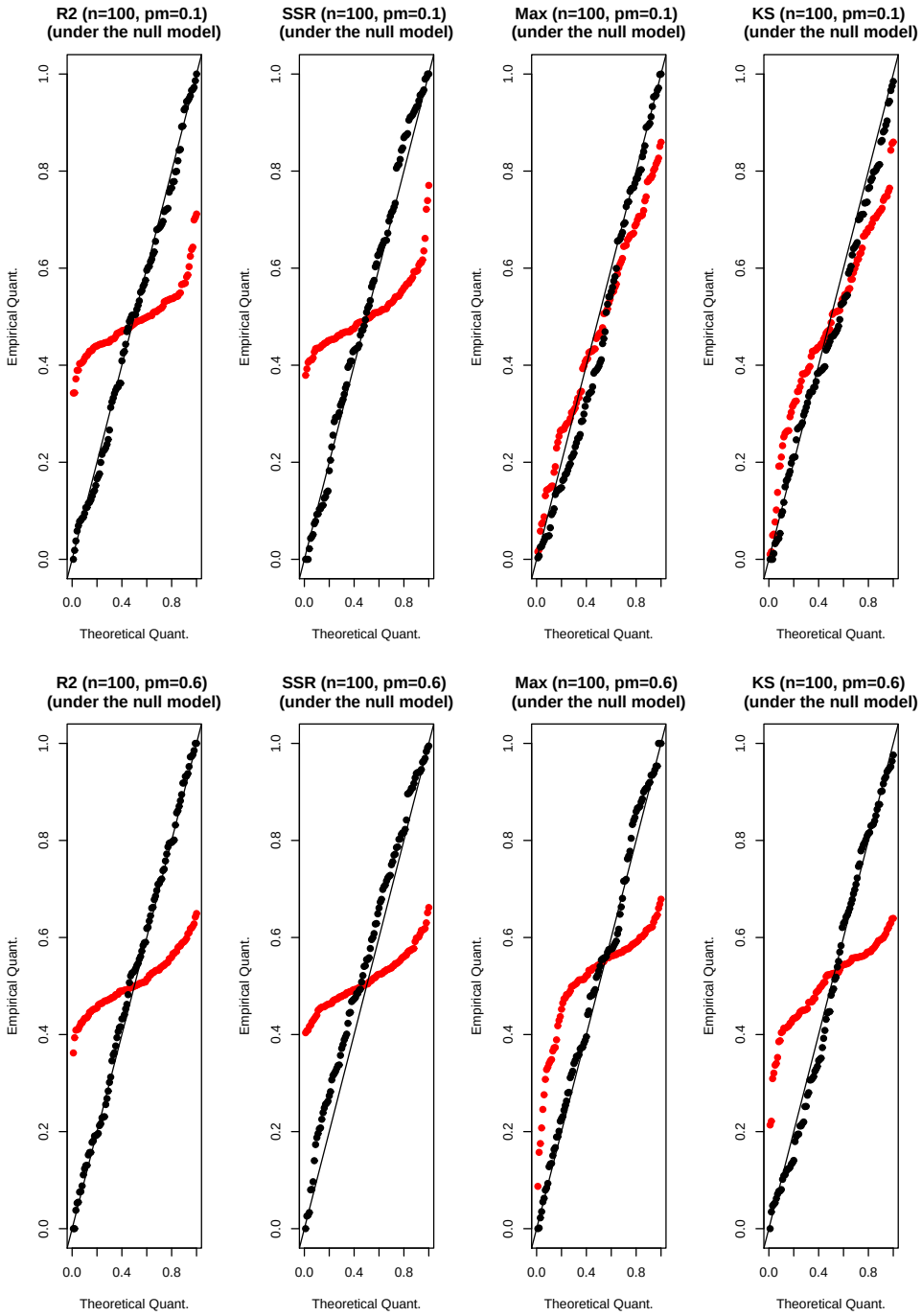

Figure 1: For the case of an improper prior and a minimum training sample, this figure shows, through QQ-plots, sampling distributions of *cppp* (black) and *ppp* (red), under the null model, where missing values are only in $\boldsymbol{y}$, with $n = 100$ and $n = 500$ and two proportions of missing values, namely 10% and 60%. Posterior predictive p-values are not uniformly distributed under the null hypothesis.

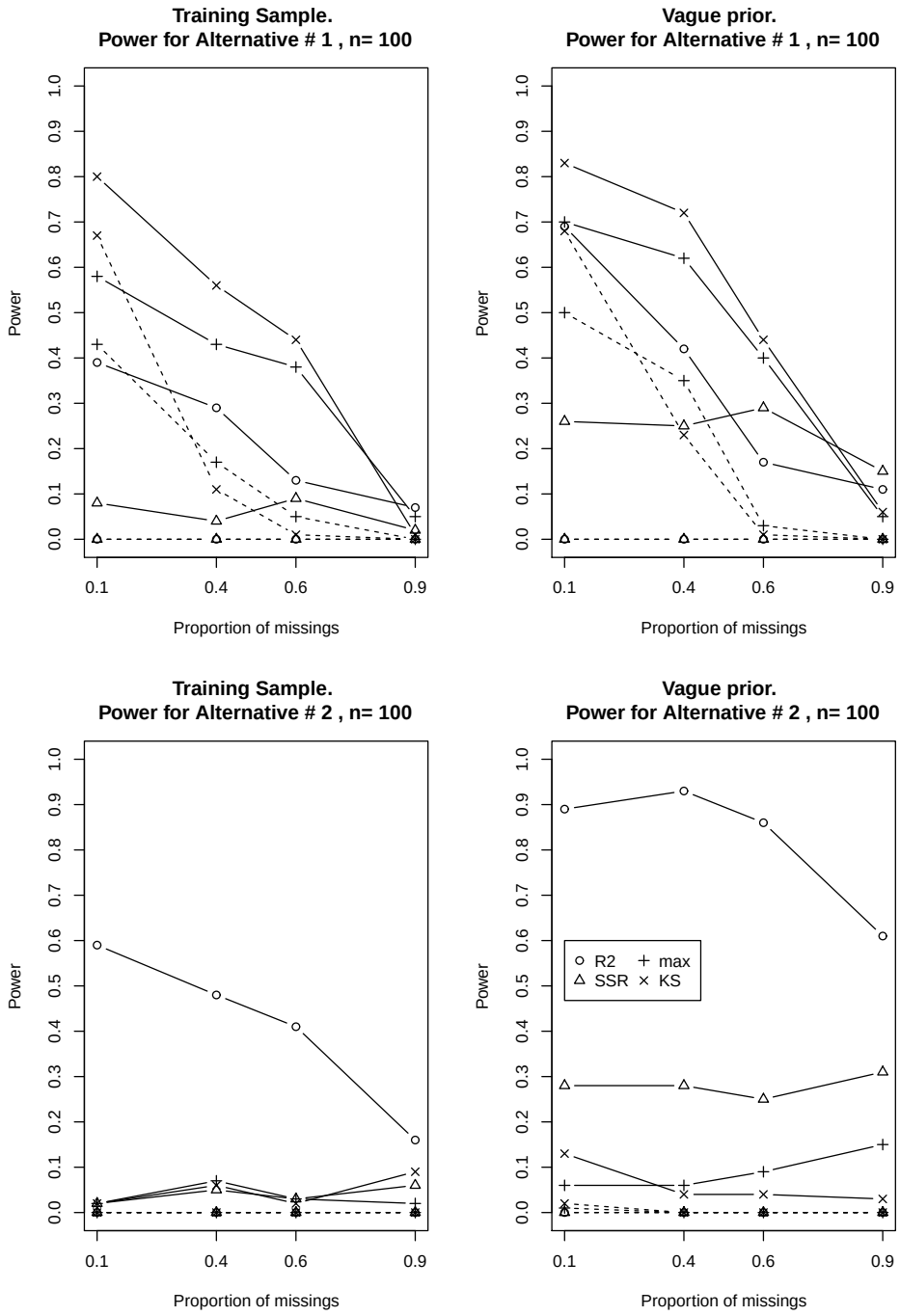

Figure 2: Powers of *cppp* (continuous lines) and *ppp* (dashed lines), under rejection of the null with p-value less than 0.05, and missing values only in $\boldsymbol{y}$. Powers are calculated with $n = 100$ and proportion of missing varying from 10% up to 90%.

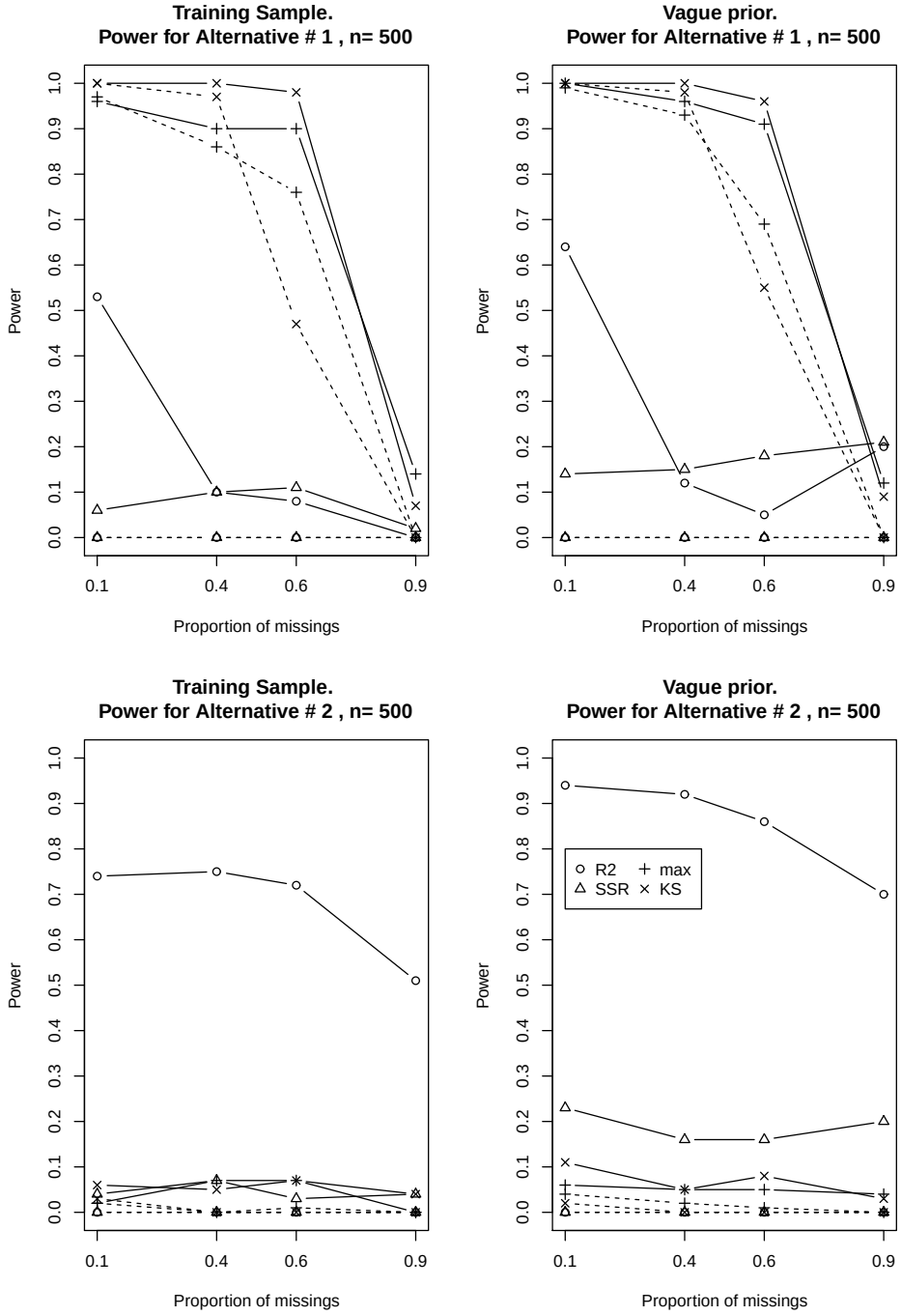

Figure 3: Powers of *cppp* (continuous lines) and *ppp* (dashed lines), under rejection of the null with p-value less than 0.05, and missing values only in $\boldsymbol{y}$. Powers are calculated with $n = 500$ and proportions of missing varying from 10% up to 90%.

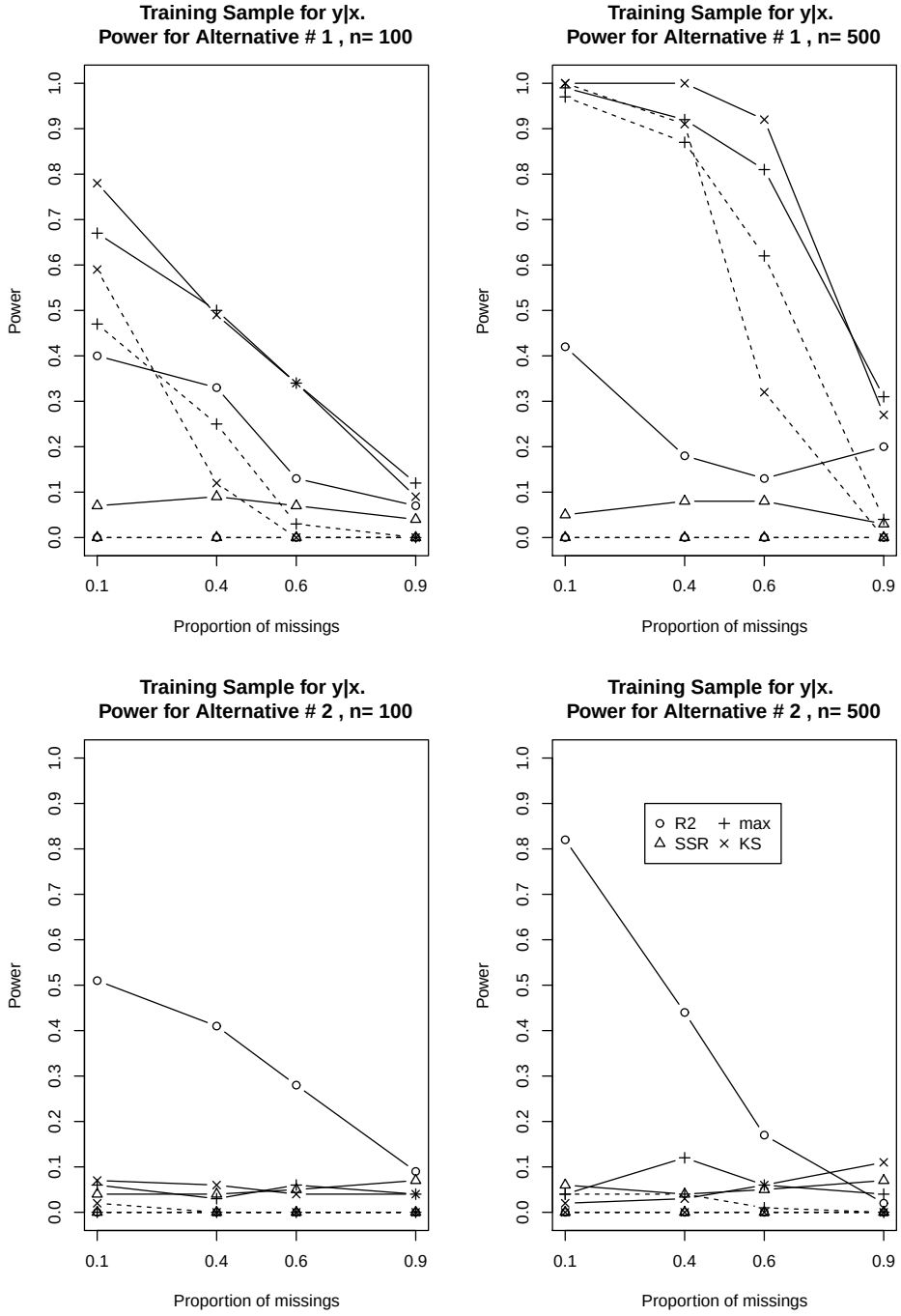

Figure 4: Powers of *cppp* (continuous lines) and *ppp* (dashed lines), for the imputation model of $\boldsymbol{y}|\boldsymbol{x}$ when rejecting the null for p-value less than 0.05. Powers are calculated with $n = 100$, $n = 500$ and proportion of missing varying from 10% up to 90% on each variable.

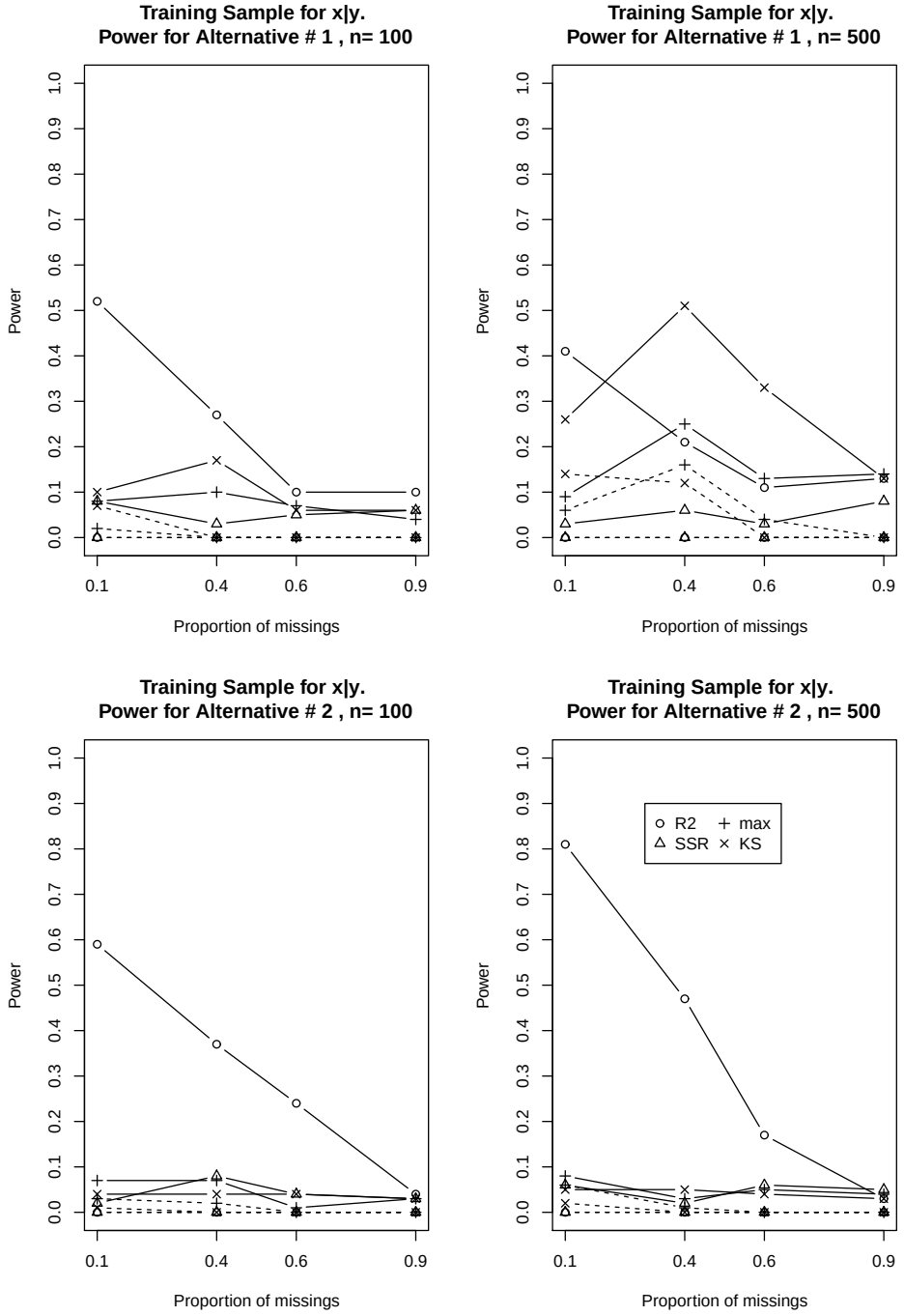

Figure 5: Powers of *cppp* (continuous lines) and *ppp* (dashed lines), for the imputation model of $\boldsymbol{x}|\boldsymbol{y}$ when rejecting the null for p-value less than 0.05. Powers are calculated with $n = 100$, $n = 500$ and proportion of missing varying from 10% up to 90% on each variable.

Finally, for the considered different degrees of freedom, n_0, in $\pi^c(\boldsymbol{\beta})$, we obtain similar results (not shown here).

It is interesting to analyze in more detail the behavior of some of these measures under both alternatives. As we have seen the calibration has no power when using $D = SSR$. This is explained, in Figures 6 and 7, by the fact that posterior distributions of $D(\boldsymbol{Y}^{\text{rep}}, \boldsymbol{\beta})|\boldsymbol{y}$ and $D((\boldsymbol{y}, \boldsymbol{Y}_m), \boldsymbol{\beta})|\boldsymbol{y}$, used to calculate *ppp* in equation (6), do not change for $\boldsymbol{y} = \boldsymbol{y}_o$ and $\boldsymbol{y} = \boldsymbol{Y}_{o,k}$ coming from alternatives 1 or 2 and the prior predictive distribution, respectively. This is due to the fact that the sampling distribution of SSR do not change for the three considered scenarios and it makes this discrepancy useless for GOF even using *cppp*.

For $D = R^2$ we observe that posterior distributions of $D(\boldsymbol{Y}^{\text{rep}}, \boldsymbol{\beta})|\boldsymbol{y}_o$ and $D((\boldsymbol{y}_o, \boldsymbol{Y}_m), \boldsymbol{\beta})|\boldsymbol{y}_o$ are located in the same region (around 0.7) for $\boldsymbol{y}_o$ coming from Alternative 1, bottom-left plot of Figure 6, while $D(\boldsymbol{Y}^{\text{rep}}, \boldsymbol{\beta})|\boldsymbol{Y}_{o,k}$ and $D((\boldsymbol{y}_o, \boldsymbol{Y}_m), \boldsymbol{\beta})|\boldsymbol{Y}_{o,k}$ are located in another region (around 0.95). In both cases the shape of the clouds is the same around the bisector, leading to the same *ppp* and rendering useless the calibration with *cppp*. This is why we cannot detect incompatibilities working with R^2 under Alternative 1. Instead, under Alternative 2, $D(\boldsymbol{Y}^{\text{rep}}, \boldsymbol{\beta})|\boldsymbol{y}_o$ is centered around 0 meanwhile $D((\boldsymbol{y}_o, \boldsymbol{Y}_m), \boldsymbol{\beta})|\boldsymbol{y}_o$ is located in smaller values, bottom-left plot in Figure 7. Also in this case, *ppp* (0.322) is not able to detect incompatibilities whereas, if we consider the calibration, we have that values $ppp(\boldsymbol{Y}_{o,k})$ for $k = 1, \ldots, K$ are around 0.5 and these values are sufficiently different from the observed one (0.322) to detect incompatibility under Alternative 2.

This simulation study is limited in the sense that it only considers two types of departures from the null model, which are the most common found in practice. Other types of departure from the null model may need other discrepancies to be detected, but a comprehensive study of such departures and discrepancies is beyond the scope of the paper. Instead, the focus of this work is to analyze the behavior of *cppp*, selecting several discrepancy measures but without an extensive study of these.

4 Application

For illustration purposes we present an application to the *boys* data set included in the `MICE` package. This data set consists of a random sample of 10% of observations from the cross-sectional data used to construct the Dutch growth references in 1997 (Fredriks et al. 2000). In this data set there are several variables related to $n = 748$ Dutch boys, from which we restrict to the following continuous ones: Age, $Height$, $Weight$ and Head Circumference (HC). A matrix of dispersion plots for these variables along with their marginal distributions appears in Figure 8. We can see that, as growth rates differ from younger to older boys, there are highly non linear relations between these four variables across all ages. In contrast, these are very well approximated by linear relations for small age groups, such as that of boys under 1 year old. Based on this, we expect that the linear imputation model (3) is compatible with data if applied to a specific age group rather than to the whole data set. The percentage of missing

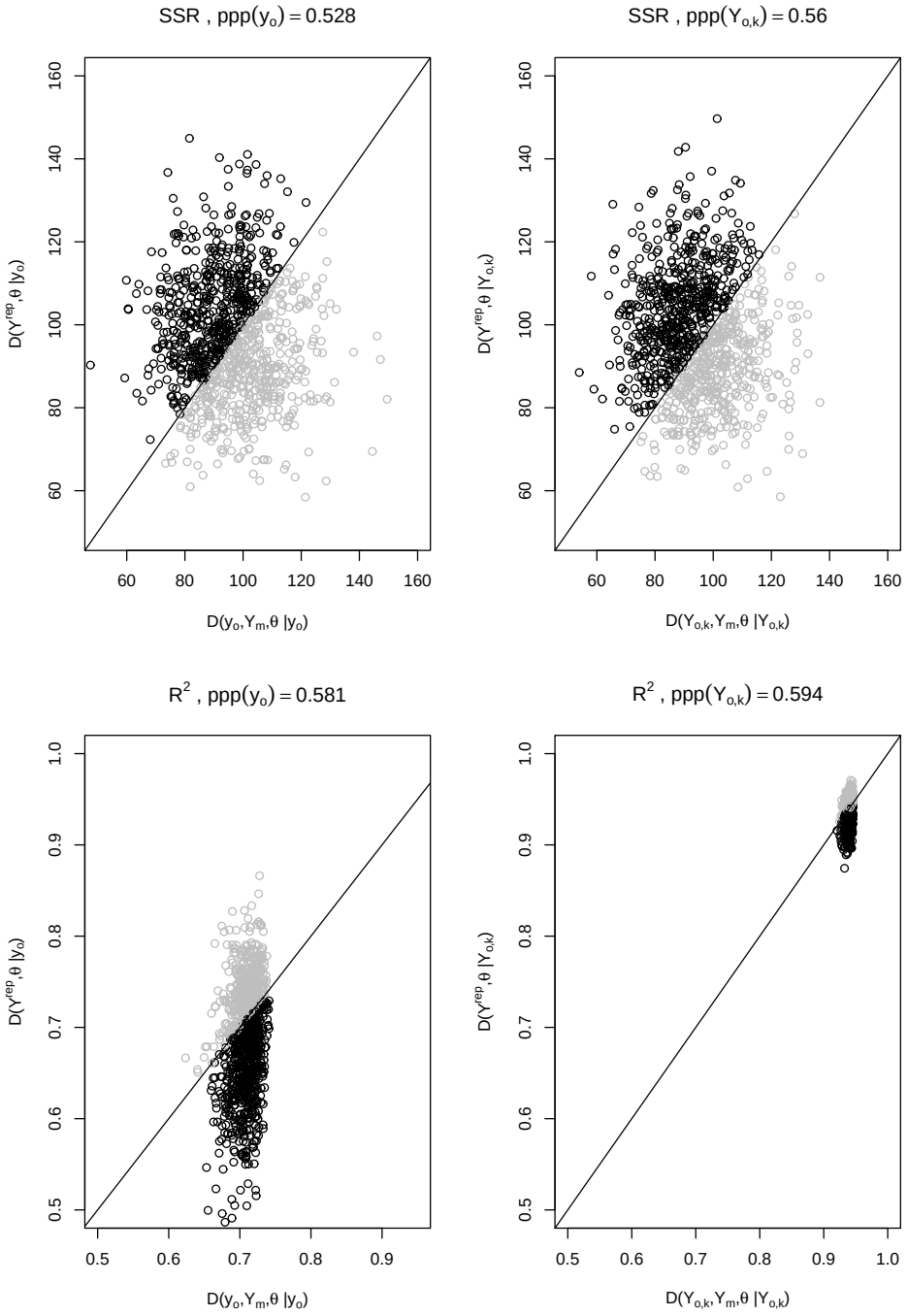

Figure 6: For a random sample of size $n = 100$ and $pm = 0.1$, generated under Alternative 1, we show, 1000 generations of $D(\boldsymbol{Y}^{\text{rep}}, \boldsymbol{\beta})|\boldsymbol{y}_o$ and $D((\boldsymbol{y}_o, \boldsymbol{Y}_m), \boldsymbol{\beta})|\boldsymbol{y}_o$ (left plots) used in (6) for SSR (top) and R^2 (bottom). In each case ppp is obtained by the proportion of black points for each discrepancy. Right plots contain 1000 generations of $D(\boldsymbol{Y}^{\text{rep}}, \boldsymbol{\beta})|\boldsymbol{Y}_{o,k}$ and $D((\boldsymbol{Y}_{o,k}, \boldsymbol{Y}_m), \boldsymbol{\beta})|\boldsymbol{Y}_{o,k}$, $\boldsymbol{Y}_{o,k} \sim m(\boldsymbol{Y})$ based on $\pi^t(\boldsymbol{\beta})$. These simulations are used to approximate $ppp(\boldsymbol{Y}_{o,k})$ in (7).

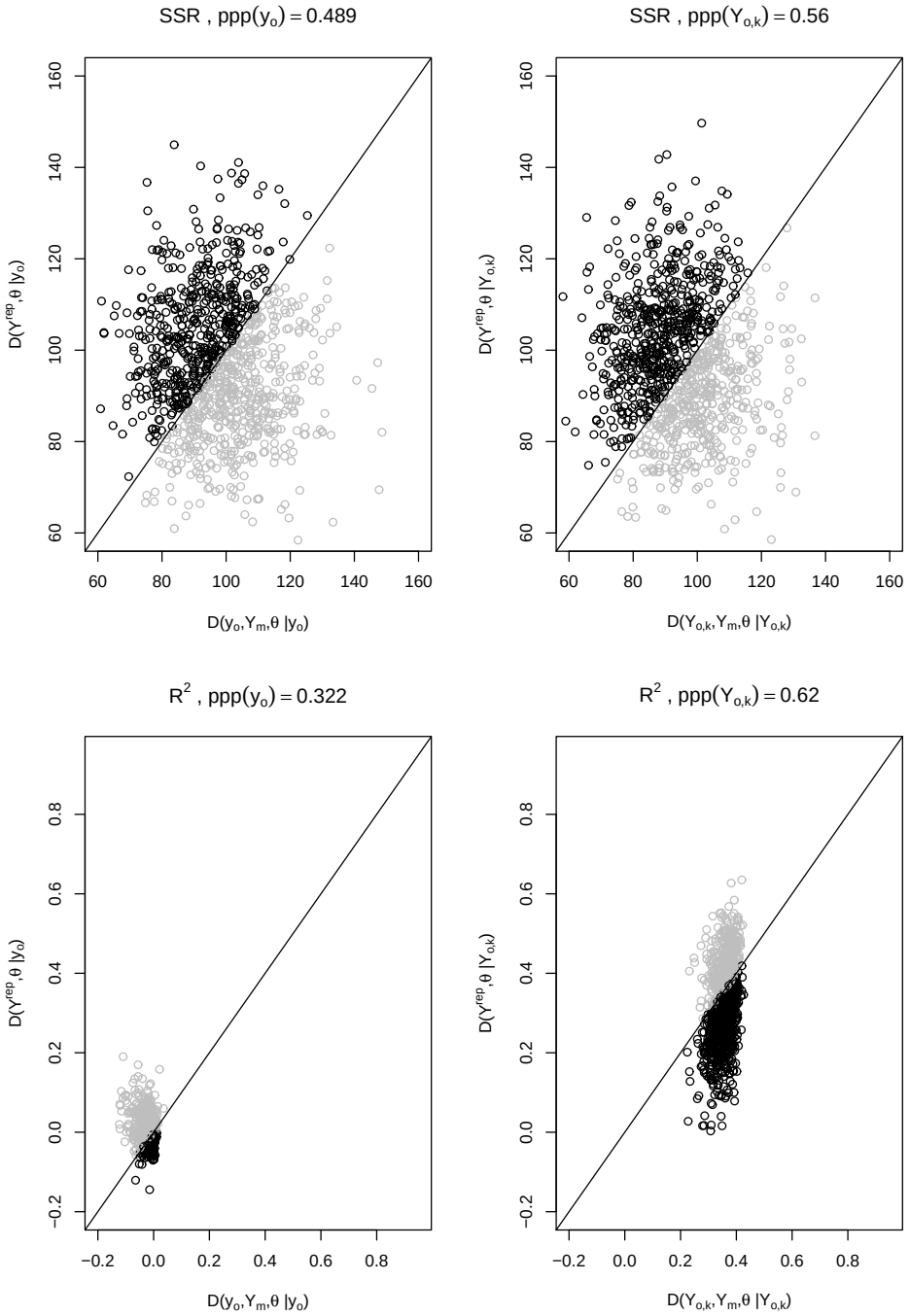

Figure 7: For a random sample of size $n = 100$ and $pm = 0.1$, generated under Alternative 2, we show, 1000 generations of $D(\boldsymbol{Y}^{\text{rep}}, \boldsymbol{\beta})|\boldsymbol{y}_o$ and $D((\boldsymbol{y}_o, \boldsymbol{Y}_m), \boldsymbol{\beta})|\boldsymbol{y}_o$ (left plots) used in (6) for $D = SSR$ (top) and $D = R^2$ (bottom). In each case *ppp* is obtained by the proportion of black points for each discrepancy. Right plots contain 1000 generations of $D(\boldsymbol{Y}^{\text{rep}}, \boldsymbol{\beta})|\boldsymbol{Y}_{o,k}$ and $D((\boldsymbol{Y}_{o,k}, \boldsymbol{Y}_m), \boldsymbol{\beta})|\boldsymbol{Y}_{o,k}$, $\boldsymbol{Y}_{o,k} \sim m(\boldsymbol{Y})$ based on $\pi^t(\boldsymbol{\beta})$. These simulations are used to approximate $ppp(\boldsymbol{Y}_{o,k})$ in (7).

values for the considered variables are: 0.5% for $Weight$; 5% for $Height$; 6% for HC; and Age is fully observed.

SRMI was iterated $S = 100$ times and we computed $\{cppp^{s,q},\ q = 1, 2, 3,\quad s = 1, \ldots, 100\}$ for assessing the GOF of each imputation model (3), where each variable was imputed given the rest, i.e. all the variables except the one being imputed are covariates in the imputation model. We considered the two classes of priors proposed here. For the conjugate prior, we used the same values of hyperparameters as in the simulation study but with $V_0^{-1} = 200\boldsymbol{I}$. As the results for both priors were similar and based on the previous simulation study, we only show the results for the trained prior $\pi^t(\boldsymbol{\beta})$, provided that n is large.

The corresponding *cppp*, for all iterations and all three imputation models, are shown in Figure 9 for Max and KS with individuals of all ages (first row). The values of *cppp* in Figure 9 suggest that the conditional regression models are incompatible with the observed data. There are problems with the normality of residuals for the three models, as shown by the *cppp* associated with the KS discrepancy. Also the Max discrepancy reflects incompatibility between the observed values and the predicted maximum for models: $Weight|rest$ and $HC|rest$. Based on this, and also on what has been noted above, we consider imputing only data for boys younger than 1 year. The corresponding *cppp* for this case are shown in the second row of Figure 9. We can see that, for this age group, all linear imputation models are compatible with the observed data. These results are in line with the above considerations on the growth rates. The same analysis was performed using *ppp* as shown in Figure 10. Results from Figures 9-12 are further summarized in Table 1, where the proportion of times in which *cppp* or *ppp* are lower than 0.05 and 0.1 are reported. These proportions approximate the power of the method when considering boys of all ages. For the case of younger boys, these quantities can be interpreted as the Type I error. Although the results for *ppp* are basically similar to those for *cppp*, it can be seen in Table 1 that power of *ppp* is 0.27 (for a cut-off of 0.05) for model $HC|rest$ when using KS in all ages, while the corresponding *cppp* are clearly below 0.05 (power 0.77).

In order to check whether the performance of the *cppp* is adequate with a higher percentage of missing values, and only for illustration purposes, we repeat the previous analysis by considering the same individuals, in which a total of 30% of the observations of each variable $Weight$, $Height$ and HC have been uniformly deleted at random. This way of introducing missing values mimics the missing completely at random mechanism, a particular case of MAR. Results of the GOF for the three imputation models over all ages and younger boys are shown in Figure 11, while corresponding *ppp* are shown in Figure 12.

Using the *cppp* we reach the same conclusions with respect to the analysis with all available data, although the *cppp* exhibits larger noise. Instead, the results obtained through the *ppp* for KS measure suggest that models $HC|rest$ and $Height|rest$ are compatible with data for all ages. These results provide evidence for the fact that the proposed method is robust with respect to the percentage of missing data.

Both Figures 9 and 11 suggest stability of the *cppp* with respect to the sequential

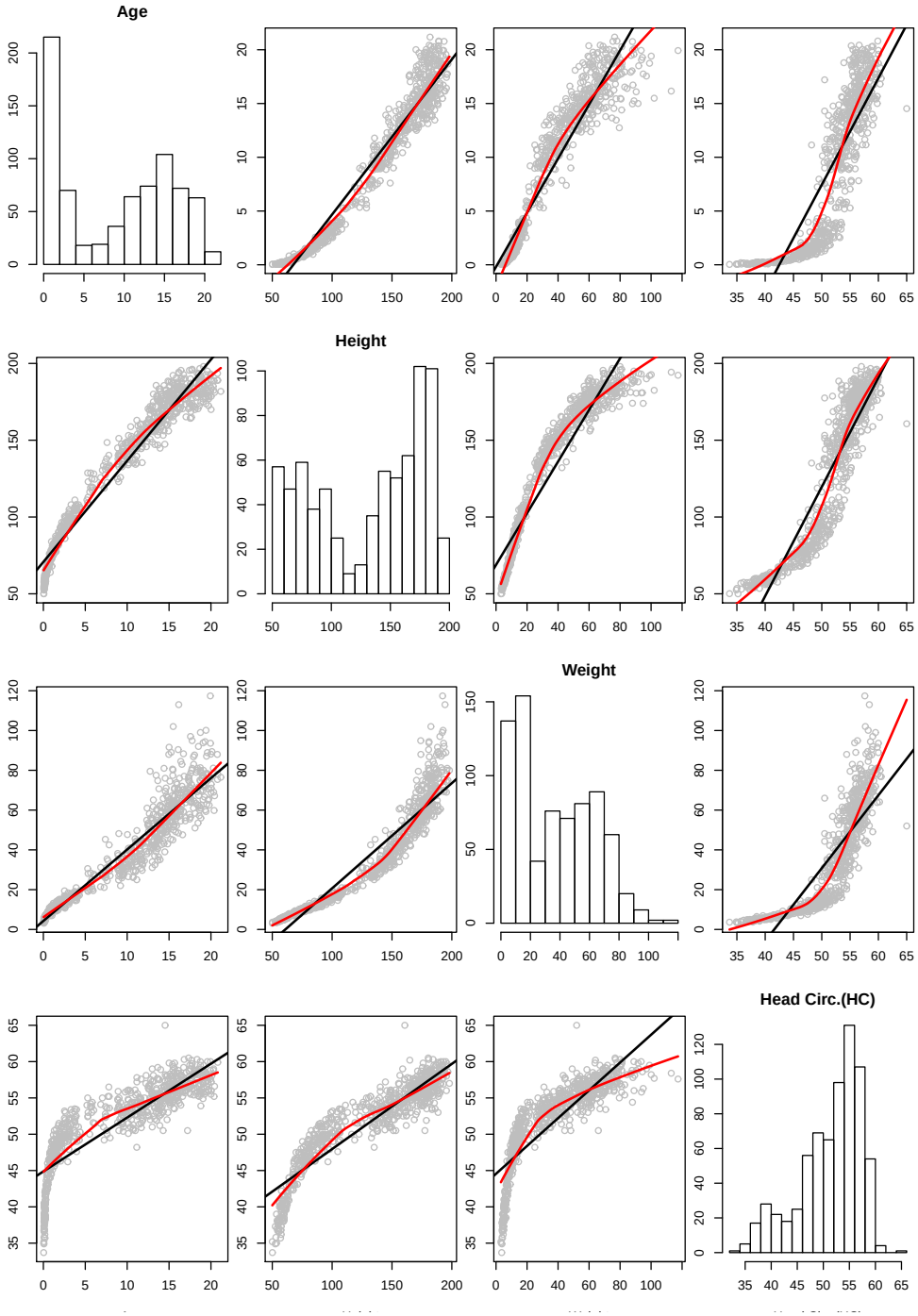

Figure 8: Boys Data. Linear Regressions (black) and Lowess Regressions (red) show that relations among variables are not linear, meanwhile for boys younger than 1 year it seems to be linear.

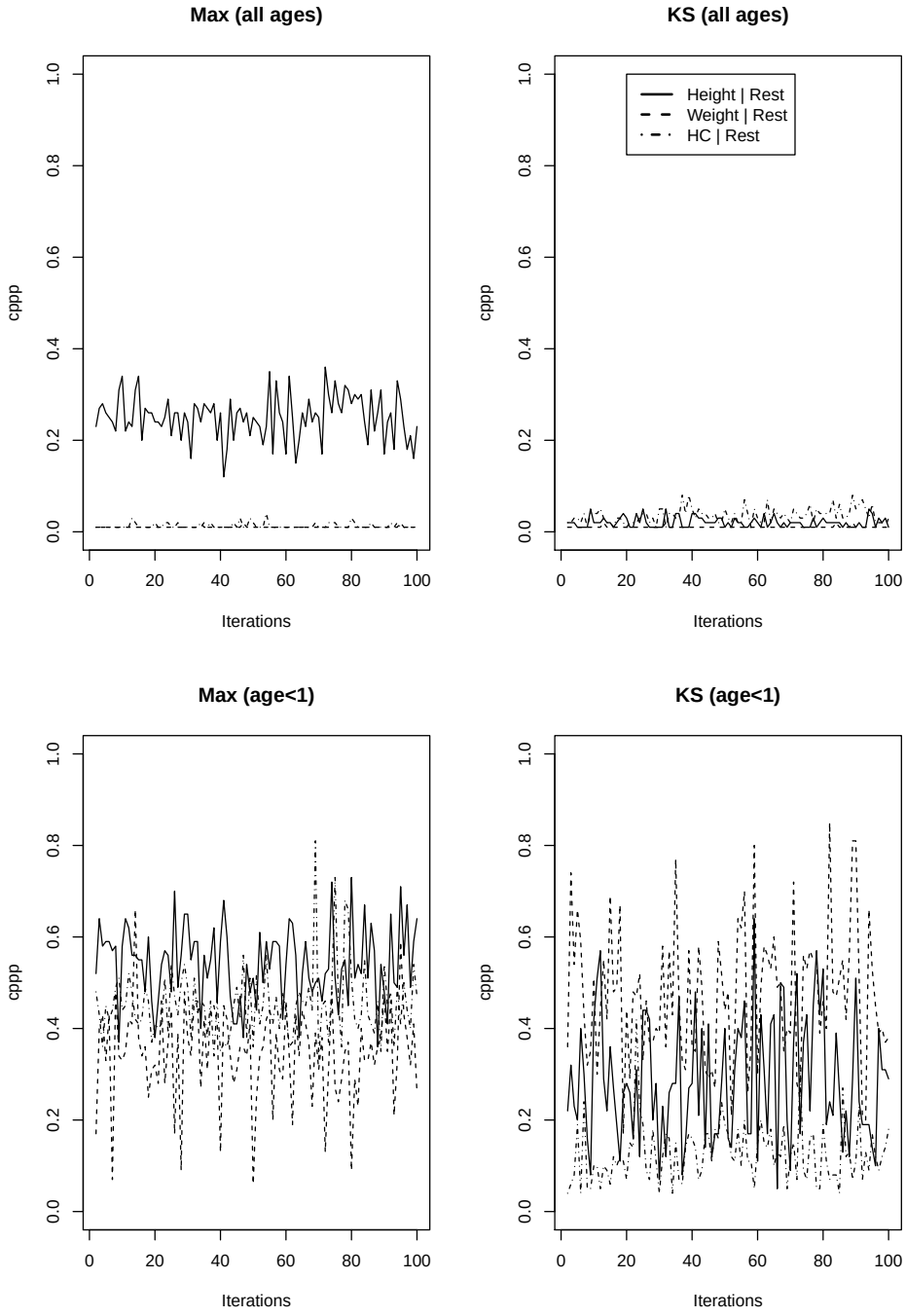

Figure 9: Calibrated posterior predictive p-values, *cppp*, for assessing the GOF of the multiple linear imputation models of a variable given the rest. We used Boys data with the original missing values. Corresponding *cppp*, for all iterations of imputation steps, are showed for Max and KS with all ages (first row) and with only younger boys (second row).

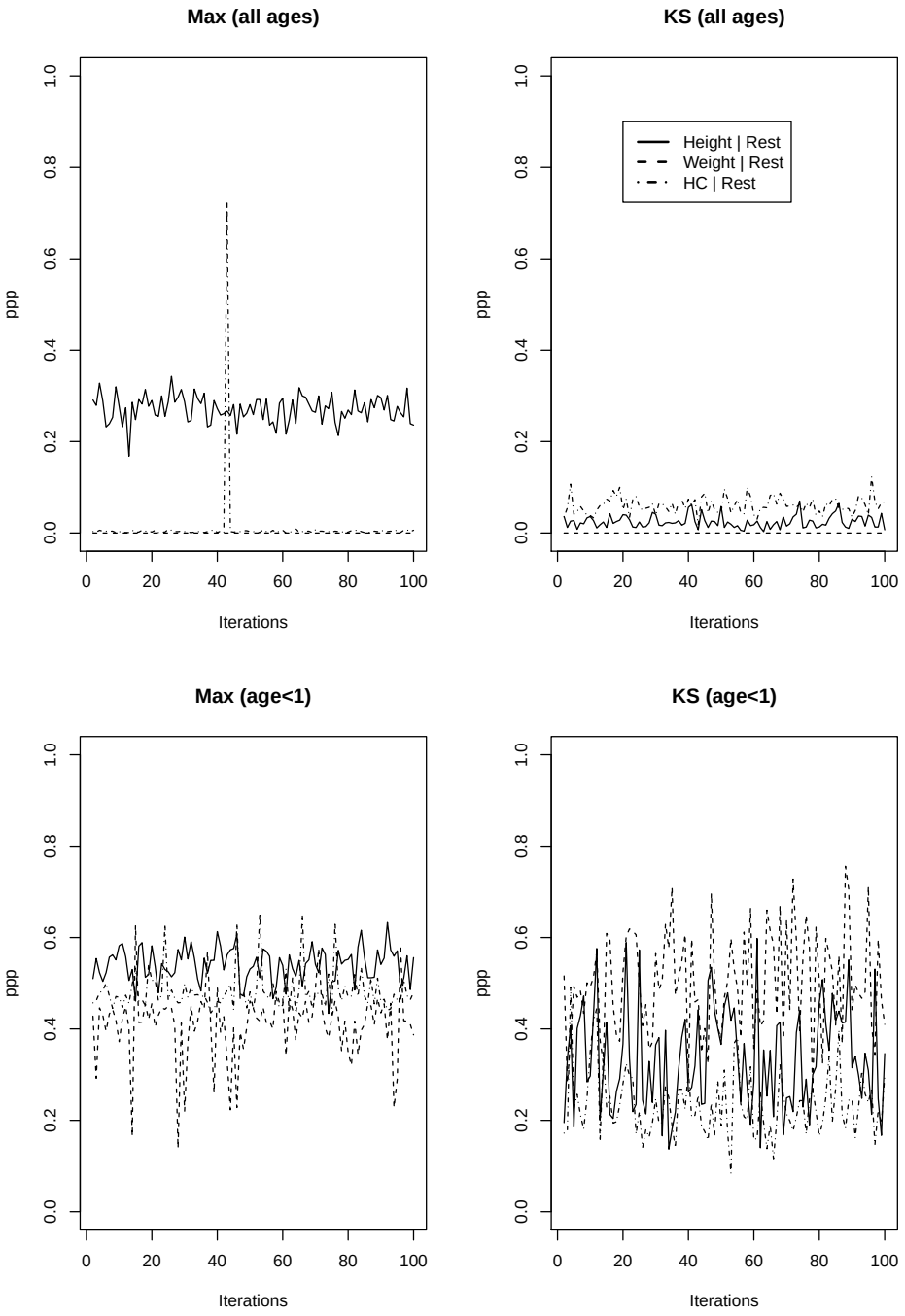

Figure 10: Posterior predictive p-values, *ppp*, for assessing the GOF of the multiple linear imputation models of a variable given the rest. We used Boys data with the original missing values. Corresponding *ppp*, for all iterations of imputation steps, are showed for Max and KS with all ages (first row) and with only younger boys (second row).

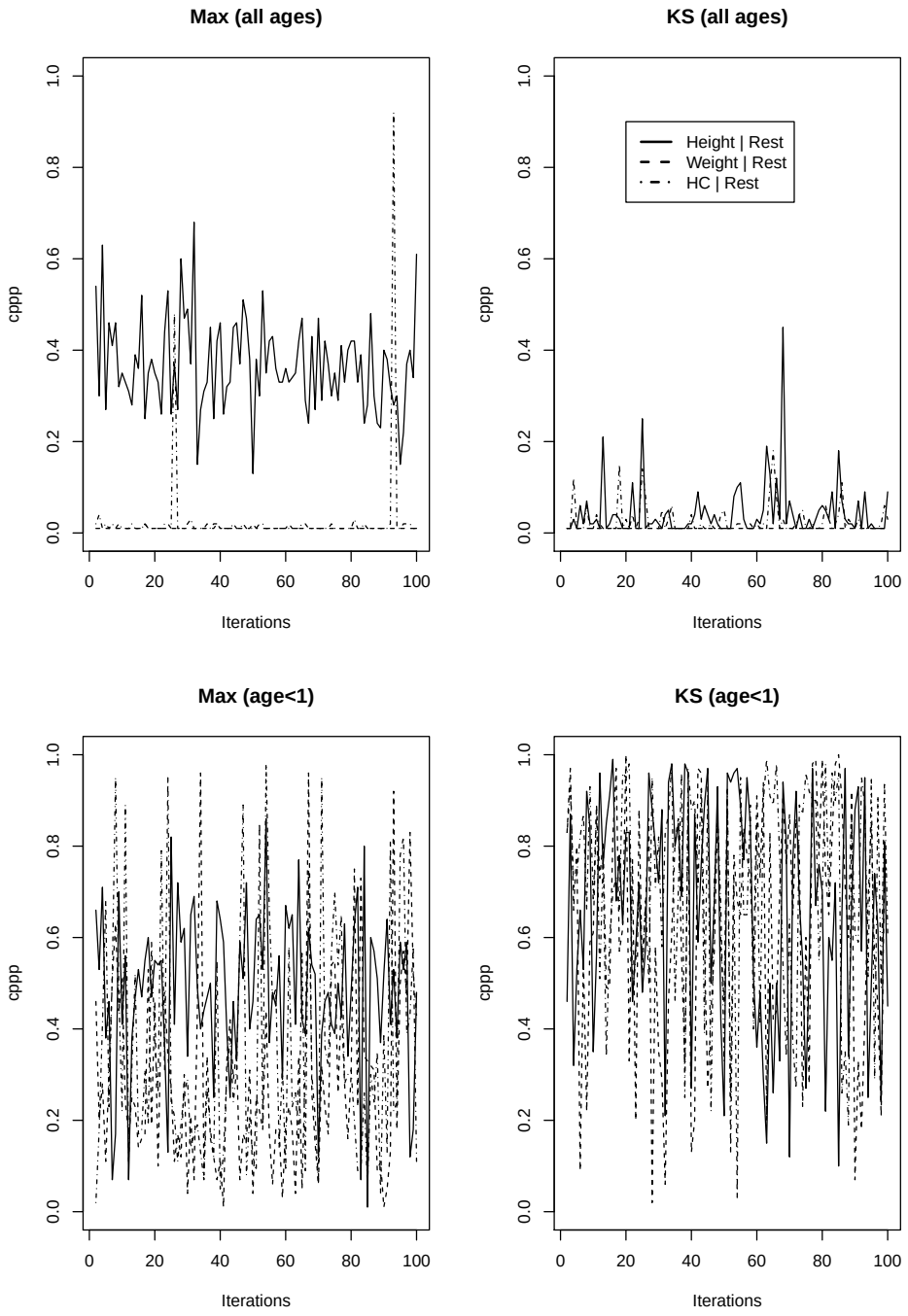

Figure 11: Calibrated posterior predictive p-values, *cppp* for assessing the GOF of the multiple linear imputation models of a variable given the rest. We used Boys data with artificial 30% of missing values in each variable. Corresponding *cppp*, for all iterations of imputation steps, are showed for Max and KS with all ages (first row) and with only younger boys (second row).

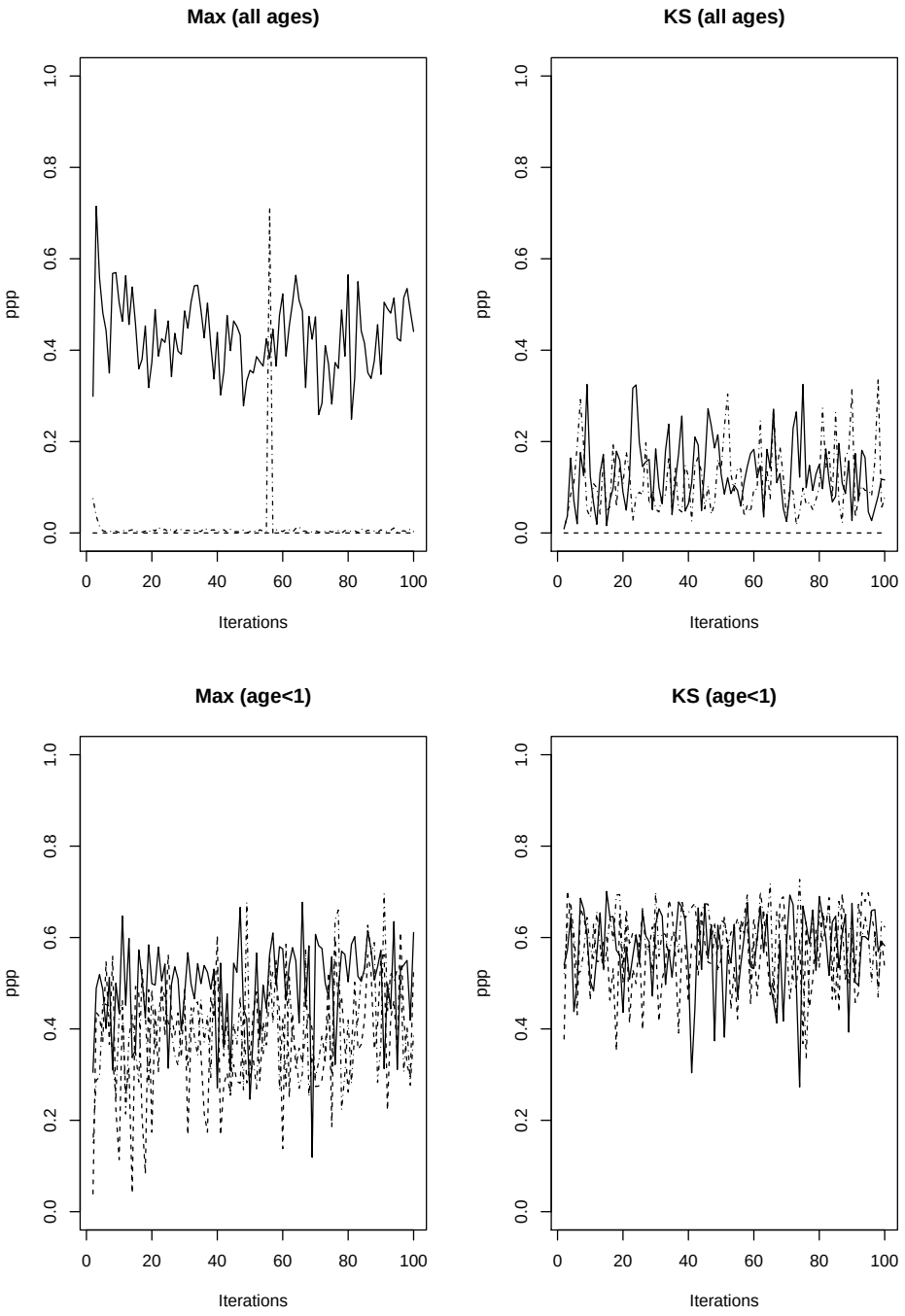

Figure 12: Posterior predictive p-values, *ppp*, for assessing the GOF of the multiple linear imputation models of a variable given the rest. We used Boys data with artificial 30% of missing values in each variable. Corresponding *ppp*, for all iterations of imputation steps, are showed for Max and KS with all ages (first row) and with only younger boys (second row).

Results with original data								
	$cppp < 0.05$				$cppp < 0.1$			
	Max		KS		Max		KS	
	all	< 1	all	< 1	all	< 1	all	< 1
Height	0.00	0.00	0.96	0.00	0.00	0.00	1.00	0.05
Weight	1.00	0.00	1.00	0.00	1.00	0.04	1.00	0.00
HC	1.00	0.00	0.77	0.05	1.00	0.00	1.00	0.33
	$ppp < 0.05$				$ppp < 0.1$			
	Max		KS		Max		KS	
	all	< 1	all	< 1	all	< 1	all	< 1
Height	0.00	0.00	0.94	0.00	0.00	0.00	1.00	0.00
Weight	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
HC	0.99	0.00	0.27	0.00	0.99	0.00	0.96	0.01

Results with 30% of missings								
	$cppp < 0.05$				$cppp < 0.1$			
	Max		KS		Max		KS	
	all	< 1	all	< 1	all	< 1	all	< 1
Height	0.00	0.01	0.74	0.00	0.00	0.04	0.90	0.00
Weight	1.00	0.06	1.00	0.01	1.00	0.15	1.00	0.04
HC	0.98	0.02	0.85	0.01	0.98	0.12	0.94	0.01
	$ppp < 0.05$				$ppp < 0.1$			
	Max		KS		Max		KS	
	all	< 1	all	< 1	all	< 1	all	< 1
Height	0.00	0.00	0.13	0.00	0.00	0.00	0.37	0.00
Weight	1.00	0.02	1.00	0.00	1.00	0.03	1.00	0.00
HC	0.98	0.00	0.18	0.00	0.99	0.00	0.55	0.00

Table 1: Table entry is the proportion that *cppp* or *ppp* is under the specified threshold. Values are obtained from Figures 9-12.

imputation, as the *cppp* does not show any trend along iterations as a consequence of the sequential imputation. The peaks, for checking $HC|Rest$ using maximum in all ages (Figures 10-12), appear because the apparent outlier in the HC measure (Figure 8) has been sampled in the training, and it is not used to calculate the discrepancy. However, this behavior is not a problem in our procedure as we change the training sample in each iteration. The noise observed in the *cppp* sequence is similar for both priors, meaning that it is due to the change in the imputed values along the imputation steps. In fact, for a percentage of missing tending to one we expect that the distribution of the *cppp* will converge to the uniform, as most of data are simulated according to the model.

Results for R^2 and SSR are not shown because the corresponding *cppp* are essentially uniformly distributed, not allowing us to assess the fit of the imputation models. This is consistent with the low power of R^2 and SSR in the simulation study, because the underlying model, for all data, is similar to that in the Alternative 1.

5 Conclusion

In this paper we show that GOF of a conditional regression imputation model, using a specific discrepancy measure, needs a calibrated measure such as *cppp*. In fact, in the simulation study, *ppp* does not exhibit a satisfactory behavior in the considered alternatives because of its lack of calibration. We propose how to employ *cppp* in the presence of missing values when assessing the GOF of SRMI.

In the simulation study we reach the conclusion that some Ds are more sensitive than others for detecting some kinds of incompatibility between the imputation null model and the observed data. Based on this, we recommend considering several discrepancy measures when assessing the fit of SRMI, as illustrated in the application to the Boys data set. When investigating the fit of SRMI in other types of models, it is convenient to use discrepancies or statistics more related to the assumed model. For example, in the case of general linear models, regression coefficients, their standard deviation, percentiles of complete data under the model, etc. could be used as discrepancy measures. Again, the investigation of the best discrepancy to be used is beyond the scope of this paper.

In this paper, we consider two types of priors to calibrate *ppp*, highlighting the benefits of using each of them. The use of the trained prior based on a default prior provides an automatic procedure and avoids eliciting prior parameters. On the other hand, when it is possible to elicit a prior distribution, such as the conjugate one in Section 2.2, it results in larger power to detect incompatibilities between model and prior with respect to observed data. Differences between the behavior of *cppp*, based on both priors, seem to vanish for larger sample sizes. Given the above considerations and the fact that MI is made prior to the final analysis, we recommend the use of the trained prior approach.

The final message is that the *cppp* has a better performance than the *ppp* when assessing the GOF of regression imputation models. However the *cppp* depends on the prior distribution specified and on the behavior of the *ppp* for a given discrepancy as

illustrated in Section 3 and in Figures 6 and 7. There are other measures to assess the GOF that are uniformly distributed under the null and whose investigation for SRMI is beyond the scope of this paper. For example, we may employ the prior predictive p-value, advocated by Box (1980), in which $m(\boldsymbol{Y})$ induces a sampling distribution of a discrepancy measure under the null model for a given proper prior, $\pi(\boldsymbol{\beta})$. Another proposal would be the simulation-based model checking in Dey et al. (1998), which also requires a proper prior distribution. It would be of interest to investigate the behavior of such measures when using the trained prior proposed in this work. Further possibilities would be the conditional and partial predictive p-values, proposed in Bayarri and Berger (1999, 2000), applied to validate the GOF of several models in Bayarri and Castellanos (2001, 2007). However, in order to employ such measures, it is necessary to work with statistics which is somewhat less appropriate for GOF of imputation models as this approach implies using only observed data.

Minimum training samples (MTS) are drawn using an equiprobable distribution in the observed data $\boldsymbol{y}_t$, as explained in Section 2.1. This strategy of MTS simulation results in samples that are not random samples from the complete original data set, $\boldsymbol{Y} = (\boldsymbol{y}_o, \boldsymbol{Y}_m)$, as we are assuming a MAR mechanism for $\boldsymbol{Y}$. Making inference about the probability of missingness could be an interesting piece of future work in order to use these estimated probabilities to select the training sample, as discussed in Berger and Pericchi (2004). On the other hand, the use of other optimality criteria to select the MTS, as the information of each observation (Berger and Pericchi 2004), could avoid, for instance, the peaks appearing in Figures 10-12.

References

Abayomi, K., Gelman, A., and Levy, M. (2008). "Diagnostics for Multivariate Imputations." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 57(3): 273–291. 430

Bayarri, M. J. and Berger, J. O. (1999). "Quantifying surprise in the data and model verification." In Bernardo, J., Berger, J., Dawid, A., and Smith, A. (eds.), *Bayesian Statistics 6*, 53–82. Oxford University Press. 453

— (2000). "P Values for Composite Null Models." *Journal of the American Statistical Association*, 95(452): 1127–1142. 430, 453

Bayarri, M. J. and Castellanos, M. (2001). "A comparison between p-values for goodness-of-fit checking." In George, E. (ed.), *Monographs of Official Statistics. Bayesian Methods with Applications to Science, Policy and Official Statistics*, 1–10. Eurostat. 453

— (2007). "Bayesian Checking of the Second Levels of Hierarchical Models." *Statistical Science*, 22(3): 322–343. 453

Berger, J. O. and Pericchi, L. (1996). "The intrinsic Bayes Factor for Model Selection and Prediction." *Journal of the American Statistical Association*, 91(433): 109–122. 431

Berger, J. O. and Pericchi, L. (2004). "Training Samples in Objective Bayesian Model Selection." *The Annals of Statistics*, 32(3): 849–869. 453

Box, G. (1980). "Sampling and Bayes inference in scientific modeling and robustness." *Journal of the Royal Statistical Society: Series A*, 143: 383–430. 453

Dahl, F. (2006). "On the conservativeness of posterior predictive p-values." *Statistics and Probability Letters*, 76: 1170–1174. 430

Dey, D., Gelfand, A., Swartz, T., and Vlachos, P. (1998). "A simulation-intensive approach for checking hierarchical models." *Test*, 7: 325–346. 453

Fredriks, A., van Buuren, S., Burgmeijer, R., Meulmeester, J., Beuker, R., Brugman, E., Roede, M., Verloove-Vanhorick, S., and Wit, J. (2000). "Continuing positive secular growth change in The Netherlands 1955-1997." *Pediatric Research*, 47: 216–323. 442

Gelman, A. (2004). "Exploratory data analysis for complex models." *Journal of Computational and Graphical Statistics*, 13: 755–787. 430

Gelman, A., Mechelen, I. V., and Verbeke, G. (2005). "Multiple Imputation for Model Checking: Complete-Data Plots with Missing and Latent Data." *Biometrics*, 61: 74–85. 430

Gelman, A., Meng, X., and Stern, H. (1996). "Posterior predictive assessment of model fitness via realized discrepancies with discussion." *Statistica Sinica*, 6: 733–807. 430

Gelman, A. and Speed, T. (1993). "Characterizing a joint probability distribution by conditionals." *Journal of the Royal Stastical Society, B*, 55: 185–188. 434

He, Y., Zaslavsky, A. M., Harrington, D. P., Catalano, P., and Landrum, M. B. (2007). "Imputation in a Multiformat and Multiwave Survey of Cancer Care." In *Proceedings of the Survey Research methods section*. American Statistical Association, 1541-1549. `www.amstat.org/sections/srms/proceedings/` 430

Hjort, N. L., Dahl, F. A., and Steinbakk, G. H. (2006). "Post-Processing Posterior Predictive p Values." *Journal of the American Statistical Association*, 101(475): 1157–1174. 430, 431

Meng, X. (1994). "Posterior predictive p-values." *Annals of Statistics*, 22: 1142–1160. 430

O'Hagan, A. (2003). "HSSS model criticism (with discussion)." In Green, P.J., Hjort, N. L. and Richardson, S. T. (eds.), *Highly Structured Stochastic Systems*, 423–445. Oxford Univ. Press. 431

Raghunathan, T., Lepkowski, J., Van Hoewyk, J., and Solenberger, P. (2001). "A multivariate technique for multiply imputing missing values using a sequence of regression models." *Survey Methodology*, 27(1): 85–95. 430, 434

Robins, J., van der Vaart, A., and Ventura, V. (2000). "The asymptotic distribution of p-values in composite null models." *Journal of the American Statistical Association*, 95: 1143–1156. 430

Rubin, D. B. (1978). "Inference and missing data." *Biometrika*, 63: 581–592. 429

— (1984). "Bayesian justifiable and relevant frequency calculations for the applied statistician." *Annals of Statistics*, 12: 1151–1172. 430

— (1996). "Multiple Imputation After 18+ Years." *Journal of the American Statistical Association*, 91(434): 473–489. 429

— (2004). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons Inc. Hoboken, New Jersey. 429

Schafer, J. L. (1999). "Multiple imputation: a primer." *Statistical Methods in Medical Research*, 8: 3–15. 429

van Buuren, S. (2007). "Multiple imputation of discrete and continuous data by fully conditional specification." *Statistical Methods in Medical Research*, 16: 219–242. 434

Acknowledgments

The authors would like to thank the referee and the Associate Editor for helpful comments. Cabras was partially supported by the Italian Ministry of Education, University and Research. M.E. Castellanos was partially supported by the Spanish Ministry of Science and Technology, under Grant MTM2010-19528 and by grant S2009/esp-1594 from CAM. Part of the work was done while M.E. Castellanos was visiting the Department of Mathematics at the University of Cagliari, Italy.