Modeling confounding by half-sibling regression

Bernhard Schölkopf[a,1], David W. Hogg[b], Dun Wang[b], Daniel Foreman-Mackey[b], Dominik Janzing[a], Carl-Johann Simon-Gabriel[a], and Jonas Peters[a]

[a]Department of Empirical Inference, MPI for Intelligent Systems, Max Planck Institute for Intelligent Systems, 72076 Tuebingen, Germany; and [b]Center for Cosmology and Particle Physics, New York University, New York, NY 10003

Edited by Richard M. Shiffrin, Indiana University, Bloomington, IN, and approved April 5, 2016 (received for review June 18, 2015)

We describe a method for removing the effect of confounders to reconstruct a latent quantity of interest. The method, referred to as "half-sibling regression," is inspired by recent work in causal inference using additive noise models. We provide a theoretical justification, discussing both independent and identically distributed as well as time series data, respectively, and illustrate the potential of the method in a challenging astronomy application.

machine learning | causal inference | astronomy | exoplanet detection | systematic error modeling

We assay a method for removing the effect of confounding noise, based on a hypothetical underlying causal structure. The method does not infer causal structures; rather, it is influenced by a recent thrust to try to understand how causal structures facilitate machine learning tasks (1).

Causal graphical models as pioneered by (2, 3) are joint probability distributions over a set of variables $X_1, \ldots, X_n$, along with directed graphs (usually, acyclicity is assumed) with vertices X_i, and arrows indicating direct causal influences. By the causal Markov assumption, each vertex X_i is independent of its nondescendants, given its parents.

There is an alternative view of causal models, which does not start from a joint distribution. Instead, it assumes a set of jointly independent noise variables, one for each vertex, and a "structural equation" for each variable that describes how the latter is computed by evaluating a deterministic function of its noise variable and its parents (2, 4, 5). This view, referred to as a functional causal model (or nonlinear structural equation model), leads to the same class of joint distributions over all variables (2, 6), and we may thus choose either representation.

The functional point of view is useful in that it often makes it easier to come up with assumptions on the causal mechanisms that are at work, i.e., on the functions associated with the variables. For instance, it was recently shown (7) that assuming nonlinear functions with additive noise renders the two-variable case identifiable (i.e., a case where conditional independence tests do not provide any information, and it was thus previously believed that it is impossible to infer the structure of the graph based on observational data).

In this work we start from the functional point of view and assume the underlying causal graph shown in Fig. 1. In the present paper, N, Q, X, Y are random variables (RVs) defined on the same underlying probability space. We do not require the ranges of the RVs to be $\mathbb{R}$, and, in particular, they may be vectorial; we use n, q, x, y as placeholders for the values these variables can take. All equalities regarding RVs should be interpreted to hold with probability one. We further (implicitly) assume the existence of conditional expectations.

Note that, although the causal motivation was helpful for our work, one can also view Fig. 1 as a directed acyclic graph (DAG) without causal interpretation, i.e., as a directed graphical model. We need Q and X (and, in some cases, also N) to be independent (denoted by $Q \perp\!\!\!\perp X$), which follows from the given structure no matter whether one views this as a causal graph or as a graphical model.

In the section *Half-Sibling Regression*, we present the method. Following this, we describe the application and provide experimental results, and conclusions. Note that the present work extends a conference presentation (8). Proofs that are contained in ref. 8 have been relegated to *Supporting Information*.

Half-Sibling Regression

Suppose we are interested in the quantity Q, but, unfortunately, we cannot observe it directly. Instead, we observe Y, which we think of as a degraded version of Q that is affected by noise N. Clearly, without knowledge of N, there is no way to recover Q. However, we assume that N also affects another observable quantity (or a collection of quantities) X. By the graph structure, conditional on Y, the variables Q and X are dependent (in the generic case); thus X contains information about Q once Y is observed (although $X \perp\!\!\!\perp Q$). This situation is quite common if X and Y are measurements performed with the same apparatus, introducing the noise N. In the physical sciences, this is often referred to as "systematics," to convey the intuition that these errors are not simply due to random fluctuations but are caused by systematic influences of the measuring device. In our application below, both types of errors occur, but we will not try to tease them apart. Our method addresses errors that affect both X and Y, for instance by acting on N, no matter whether we call them random or systematic.

We need to point out a fundamental limitation at the beginning. Even from infinite data, only partial information about Q is available, and certain degrees of freedom remain. In particular, given a reconstructed Q, we can always construct another one by applying an invertible transformation to it, and incorporating its inverse into the function computing Y from Q and N. This includes the possibility of adding an offset, which we will see below.

The intuition behind our approach is as follows. Because $X \perp\!\!\!\perp Q$, X cannot predict Q and thus Q's influence on Y. It may contain information, however, about the influence of N on Y, because X is also influenced by N. Now suppose we try to predict Y from X. As argued above, whatever comes from Q cannot be predicted; hence only the component coming from N will be picked up. Trying to predict Y from X is thus a vehicle to selectively capture N's influence on Y, with the goal of subsequently removing it, to obtain an idealized estimate of Q referred to as $\hat{Q}$.

Definition 1.

$$\hat{Q} := Y - E[Y|X] \qquad [1]$$

We first show that $\hat{Q}$ reconstructs Q (up to its expectation $E[Q]$) at least as well as $Y - E[Y]$ does.

This paper results from the Arthur M. Sackler Colloquium of the National Academy of Sciences, "Drawing Causal Inference from Big Data," held March 26–27, 2015, at the National Academies of Sciences in Washington, DC. The complete program and video recordings of most presentations are available on the NAS website at www.nasonline.org/Big-data.

Author contributions: B.S., D.W.H., D.J., and J.P. designed research; B.S., D.W.H., D.W., D.F.-M., D.J., C.-J.S.-G., and J.P. performed research; B.S., D.W.H., D.W., D.F.-M., C.-J.S.-G., and J.P. analyzed data; and B.S., D.W., D.F.-M., D.J., C.-J.S.-G., and J.P. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. Email: bs@tuebingen.mpg.de.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1511656113/-/DCSupplemental.

Proposition 1. *For any RVs Q, X, Y that satisfy $Q \perp\!\!\!\perp X$, we have*

$$E\left[(Q - E[Q] - (Y - E[Y]))^2\right] \geq E\left[\left(Q - E[Q] - \hat{Q}\right)^2\right].$$

The above also holds true if we subject Y to some transformation before substituting it into [**1**] and Proposition 1. Note that it is conceivable that our procedure might benefit from such a transformation; however, finding it is not a topic of the present paper. See also the discussion on the function g following [**2**].

Proof. We have

$$\begin{aligned} &E\left[(Q - E[Q] - (Y - E[Y]))^2\right] \\ &= E\left[(Q - Y - E[Q] + E[Y])^2\right] \\ &\geq E\left[(Q - Y - E[Q - Y|X])^2\right] \\ &= E\left[\left(Q - E[Q] - \hat{Q}\right)^2\right]. \end{aligned}$$

Note that the two terms related by the inequality in Proposition 1 differ exactly by the amount $\mathrm{Var}[E[Q - Y|X]]$. This is due to Lemma 1 (*Supporting Information*) and the law of total variance $\mathrm{Var}[Q - Y] = \mathrm{Var}[E[Q - Y|X]] + E[\mathrm{Var}[Q - Y|X]]$.

Our method is surprisingly simple, and although we have not seen it in the same form elsewhere (including standard works on error modeling, e.g., ref. 9), we do not want to claim originality for it. Related tricks are occasionally applied in practice, often using factor analysis to account for confounding effects (10–15). However, those methods usually aim at removing spurious associations between at least two variables (which could erroneously be attributed to a nonexisting causal influence of one variable on another one), whereas our method aims at reconstructing a single variable of interest. Common to these methods and ours is the idea that removal of the disturbing influence is enabled by the fact that it perturbs several or even a large number of variables at the same time, and by strong assumptions about how the disturbance acts (such as linearity or parametric models). Note that our variable X is referred to as "negative controls" in ref. 15.

We next propose an assumption that allows for a practical method to reconstruct Q up to an additive offset. The assumption allows a theoretical analysis that provides insight into why and when our method works.

Additive Systematic Error. Inspired by recent work in causal inference, we use nonlinear additive noise models (7). Specifically, we assume that there exists a function f such that

$$Y = Q + f(N). \qquad [2]$$

We do not consider the more general form $Y = g(Q) + f(N)$ because, without additional information, g would not be identifiable. Note, moreover, that, although for ref. 7, the input of f is observed and we want to decide if it is a cause of Y, in the present setting, the input of f is unobserved (16) and the goal is to recover Q, which, for ref. 7, played the role of the noise.

Complete Information. For an additive model [**2**], our intuition can be formalized: In this case, we can predict the additive component in Y coming from N, remove the confounding effect of N, and thus reconstruct Q (up to an offset)—which is exactly what we want.

Proposition 2. *Suppose N, X are RVs, and f is a measurable function. If there exists a function ψ such that*

$$f(N) = \psi(X), \qquad [3]$$

i.e., $f(N)$ can in principle be predicted from X perfectly, then we have

$$f(N) = E[f(N)|X]. \qquad [4]$$

If, moreover, the additive model assumption [**2**] *holds, with RVs Q, Y and $Q \perp\!\!\!\perp X$, then*

$$\hat{Q} = Q - E[Q]. \qquad [5]$$

In our main application below, N will be systematic errors from an astronomical spacecraft and telescope, Y will be a star under analysis, and X will be a large set of other stars. In this case, the assumption that $f(N) = \psi(X)$ has a concrete interpretation: It means that the device can be self-calibrated based on measured science data only (17) (as opposed to requiring separate calibration data).

Proposition 2 provides us with a principled recommendation on how to remove the effect of the noise and reconstruct the unobserved Q up to its expectation $E[Q]$: We need to subtract the conditional expectation (i.e., the regression) $E[Y|X]$ from the observed Y (Definition 1). The regression $E[Y|X]$ can be estimated from observations (x_i, y_i) using (linear or nonlinear) off-the-shelf methods. We refer to this procedure as "half-sibling regression" to reflect the fact that we are trying to explain aspects of the child Y by regression on its half-sibling(s) X to reconstruct properties of its unobserved parent Q.

Note that $m(x) := E[f(N)|X = x]$ is a function of x, and $E[f(N)|X]$ is the RV $m(X)$. Correspondingly, [**4**] is an equality of RVs. By assumption, all RVs live on the same underlying probability space. If we perform the associated random experiment, we obtain values for X and N, and [**4**] tells us that, if we substitute them into m and f, respectively, we get the same value with probability 1. Eq. **5** is also an equality of RVs, and the above procedure therefore not only reconstructs some properties of the unobservable RV Q—it reconstructs, up to the mean $E[Q]$, and with probability 1, the RV itself.

In practice, of course, the accuracy will depend on how well the assumptions of Proposition 2 hold. If the following conditions are met, we may expect that the procedure should work well:

i) X should be (almost) independent of Q—otherwise, our method could possibly remove parts of Q itself and thus throw out the baby with the bath water. A sufficient condition for this to be the case is that N be (almost) independent of Q, which often makes sense in practice, e.g., if N is introduced by a measuring device in a way independent of the underlying object being measured. Clearly, we can only hope to remove noise that is independent of the signal; otherwise, it would be unclear what is noise and what is signal. A sufficient condition for $N \perp\!\!\!\perp Q$, finally, is that the causal DAG in Fig. 1 correctly describes the underlying causal structure.

Note, however, that Proposition 2, and thus our method, also applies if $N \not\perp\!\!\!\perp Q$, as long as $X \perp\!\!\!\perp Q$ (see *Prediction Based on Noneffects*).

ii) The observable X is chosen such that Y can be predicted as well as possible from it; i.e., X contains enough information about

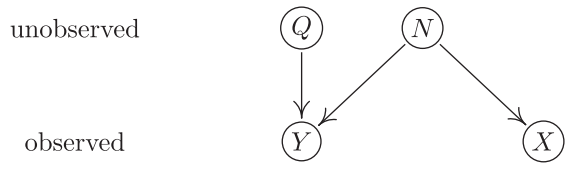

Fig. 1. We are interested in reconstructing the quantity Q based on the observables X and Y affected by noise N, using the knowledge that $(N, X) \perp\!\!\!\perp Q$. Note that the involved quantities need not be scalars, which makes the model more general than it seems at first glance. For instance, we can think of N as a multidimensional vector, some components of which affect only X, some only Y, and some both X and Y.

$f(N)$, and, ideally, N acts on both X and Y in similar ways such that a "simple" function class suffices for solving the regression problem in practice.

This may sound like a rather strong requirement, but we will see that, in our astronomy application, it is not unrealistic: X will be a large vector of pixels of other stars, and we will use them to predict a pixel Y of a star of interest. In this kind of problem, the main variability of Y will often be due to the systematic effects due to the instrument N also affecting other stars, and thus a large set of other stars will indeed allow a good prediction of the measured Y.

Note that it is not required that the underlying structural equation model be linear—N can act on X and Y in nonlinear ways, as an additive term $f(N)$.

In practice, we never observe N directly, and thus it is hard to tell whether the assumption of perfect predictability of $f(N)$ from X holds true. We now relax this assumption.

Incomplete Information. We now provide a stronger result for $\hat{Q}$ [**1**], including the case where $f(N)$ does not contain all information about X.

Proposition 3. *Let f be measurable, N, Q, X, Y be RVs with $Q \perp\!\!\!\perp X$, and $Y = Q + f(N)$. The expected squared deviation between $\hat{Q}$ and $Q - E[Q]$ satisfies*

$$E\left[\left(\hat{Q} - (Q - E[Q])\right)^2\right] = E[\mathrm{Var}[f(N)|X]]. \qquad [6]$$

Note that Proposition 2 is a special case of Proposition 3: if there exists a function ψ such that $\psi(X) = f(N)$, then the right-hand side of [**6**] vanishes. Proposition 3 drops this assumption, which is more realistic: Consider the case $X = g(N) + R$, where R is another RV. In this case, we cannot expect to reconstruct the variable $f(N)$ from X exactly.

There are, however, two settings where we would still expect good approximate recovery of Q: (*i*) If the standard deviation (SD) of R goes to zero, the signal of N in X becomes strong and we can approximately estimate $f(N)$ from X; see Proposition 4. (*ii*) Alternatively, we observe many different effects of N. In the astronomy application below, Q and R are stars, from which we get noisy observations Y and X. Proposition 5 below shows that observing many different X_i helps reconstructing Q, even if all X_i depend on N through different functions g_i and their underlying (independent) signals R_i do not follow the same distribution. The intuition is that, with increasing number of variables, the independent R_i "average" out, and thus it becomes easier to reconstruct the effect of N.

Proposition 4. *Assume that $Y = Q + f(N)$ and let*

$$X^s := g(N) + s \cdot R,$$

where R, N, and Q are jointly independent, $s \in \mathbb{R}$, $f \in C_b^0(\mathbb{R})$, g is invertible, and $g, g^{-1} \in C^0(\mathbb{R})$. Then

$$\hat{Q}^s \xrightarrow{L^2} Q - E[Q] \text{ as } s \to 0,$$

where $\hat{Q}^s := Y - E[Y|X^s]$.

Proof. We have for $s \to 0$ that

$$\begin{array}{lcl} & s \cdot R & \xrightarrow{P} 0 \\ \Rightarrow & g(N) + s \cdot R - g(N) & \xrightarrow{P} 0 \\ \overset{*}{\Rightarrow} & g^{-1}(g(N) + s \cdot R) - N & \xrightarrow{P} 0 \\ \overset{*}{\Rightarrow} & f\left(g^{-1}(g(N) + s \cdot R)\right) - f(N) & \xrightarrow{P} 0 \\ \Rightarrow & \psi_s(X^s) - f(N) & \xrightarrow{P} 0 \end{array}$$

for some ψ_s that is bounded in s (the implications $*$ follow from the continuous mapping theorem). (The notation $\xrightarrow{P}$ denotes convergence in probability with respect to the measure P of the underlying probability space.) This implies

$$E[f(N)|X^s] - f(N) \xrightarrow{L^2} 0$$

because

$$E\left[\left(f(N) - E[f(N)|X^s]\right)^2\right] \le E\left[\left(f(N) - \psi_s(X^s)\right)^2\right] \to 0$$

(L^2 convergence follows because f is bounded). However, then

$$\begin{aligned} Q - E[Q] - \hat{Q}^s &= -f(N) - E[Q] + E[f(N) + Q|X^s] \\ &= E[f(N)|X^s] - f(N) \xrightarrow{L^2} 0. \end{aligned}$$

Proposition 5. *Assume that $Y = Q + f(N)$ and that $\mathbf{X}_d := (X_1, \ldots, X_d)$ satisfies*

$$X_i := g_i(N) + R_i, \quad i = 1, \ldots, d,$$

where all R_i, N, and Q are jointly independent, $f \in C_b^0(\mathbb{R})$, $g_i \in C^0(\mathbb{R})$ for all i, $\sum_{i=1}^{\infty} \frac{1}{i^2} \mathrm{var}(R_i) < \infty$, and

$$\tilde{g}_d := \frac{1}{d} \sum_{j=1}^{d} g_j$$

is invertible with $(\tilde{g}_d^{-1})_d$ uniformly equicontinuous. Then

$$\hat{Q}_d \xrightarrow{L^2} Q - E[Q] \text{ as } d \to \infty,$$

where we define $\hat{Q}_d := Y - E[Y|\mathbf{X}_d]$.

Prediction Based on Noneffects of the Noise Variable. Although Fig. 1 shows the causal structure motivating our work, our analysis does not require a directed arrow from N to X. For the method to work in the sense of Propositions 2 and 3, we need additivity [**2**], $X \perp\!\!\!\perp Q$, and $N \not\perp\!\!\!\perp X$, to ensure that X contains information about N. We can represent this by an undirected connection between the two (Fig. 2), and note that such a dependence may arise from an arrow directed in either direction, and/or another confounder that influences both N and X. (Imagine that N is noise induced by a CCD, and X is a nearby thermometer. Both X and N will be affected by the unobserved true temperature T on the CCD and any amplifiers. The measured temperature does not have a causal effect on N, but nevertheless, it contains information about N that the regression can use.)

Knowledge about a possibly complex causal structure can help in choosing the right regression inputs X. Recall that the Markov Blanket of a node contains its parents, its children, and its children's other parents (2). Ideally, we would want X to comprise the largest possible subset of N's Markov Blanket containing information about N, subject to the constraint that $X \perp\!\!\!\perp Q$ (for an example, see Fig. S1).

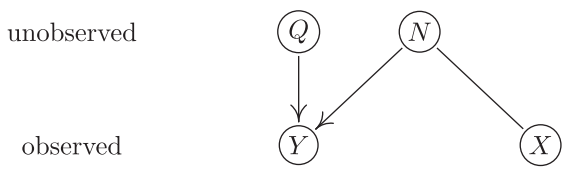

Fig. 2. Causal structure from Fig. 1 when relaxing the assumption that X is an effect of N.

Half-Sibling Regression for Time Series

Above, we have modeled the data as an independent and identically distributed (i.i.d.) sample. In practice, however, the data may be drawn from random processes that inherit a time structure. We now generalize the i.i.d. method to the time series setting.

Model Assumptions. Suppose we are given two time series, $(X_t)_t$, $(Y_t)_t$, with an ordered discrete index set, say $\mathbb{Z}$, that are generated according to the structural equation

$$\forall t \in \mathbb{Z} \quad \begin{cases} Y_t = f(N_t) + Q_t \\ X_t = g(N_t) + R_t \end{cases} \qquad [7]$$

with three unobserved time series $(Q_t)_t$, $(N_t)_t$, and $(R_t)_t$ that are independent of each other. Fig. 3 shows an example with order one processes. As before, our goal is to reconstruct the hidden time series $(Q_t)_t$.

Exploiting the Time Dependencies. Contrary to the i.i.d. case, Q_t, N_t, and R_t are now influenced by their own past Q_{t-s}, N_{t-s}, and R_{t-s}, respectively, whereas Y_t and X_t are only influenced by Q_t and N_t, respectively, as in the i.i.d. case. This may induce a time dependency in both $(X_t)_t$ and $(Y_t)_t$. We will now focus on how to exploit them to improve the reconstruction of $(Q_t)_t$.

In the i.i.d. case, we estimated Q_t with $\hat{Q}_t^S := Y_t - E[Y_t|X_t]$. In principle, we can now use the time dependency of $(X_t)_t$ by regressing Y_t on the whole future and past of $(X_t)_t$. This leads to the possibly improved estimator $\hat{Q}_t^T := Y_t - E[Y_t|(X_t)_{t\in\mathbb{Z}}]$. In some situations, it may also be useful to exploit the time dependency of $(Y_t)_t$. However, contrary to the regression of Y_t on X_t [or $(X_t)_t$], blindly regressing Y_t onto $(Y_s)_{s\neq t}$ may regress out parts of Q_t and deteriorate the results. To see this, consider Fig. 3: Although the covariates $(X_t)_t$ do not contain any information about Q_t, the covariate Y_{t-1}, for example, does; see d separation (3). Therefore, regressing Y_t on other values Y_s, $s \neq t$, can, in general, remove information of Q_t from Y_t. This may change, however, if we make additional assumptions about Q_t. This is the purpose of *Signals with Compact Support*.

Signals with Compact Support. We now assume that Q_t can be expressed as $Q_t = h(S_t, F_t)$; that is, we replace the first equation in [7] with

$$Y_t = f(N_t) + h(S_t, F_t), \qquad [8]$$

where N_{t_1}, S_{t_2}, and F_{t_3} are jointly independent for all choices of t_1, t_2, and t_3, and h is a fixed function. Denoting by $\mathcal{I}_{\text{transit}} := [t_0 - \Delta/2; t_0 + \Delta/2]$ a window of width Δ around t_0, we further assume that

$$F_t = c \text{ for } t \notin \mathcal{I}_{\text{transit}}, \qquad [9]$$

where c is a constant. In the example of exoplanet search described below, we use $h(s,f) = s \cdot f$, and $c = 1$. $(S_t)_t$ corresponds to the stellar brightness (which is variable), and the signal of interest, $(F_t)_t$, is the multiplicative change in the observed brightness due to a partial occlusion of the observed star by a planet passing through the line of sight between star and telescope. Such transits are centered around some t_0 and have a length Δ, which we think of as an RV. The RVs F_t for $t \in \mathcal{I}_{\text{transit}}$ describe the shape of the transit. Eq. **9**, however, also covers additive effects, using $h(s,f) = s + f$ and $c = 0$. We assume that t_0 is unknown and that Δ can be bounded by some α: $\mathbb{P}(\Delta < \alpha) = 1$. The goal is to detect the transits, i.e., the regions where $F_t \neq c$. We now describe the method that we will later apply to the exoplanet data set.

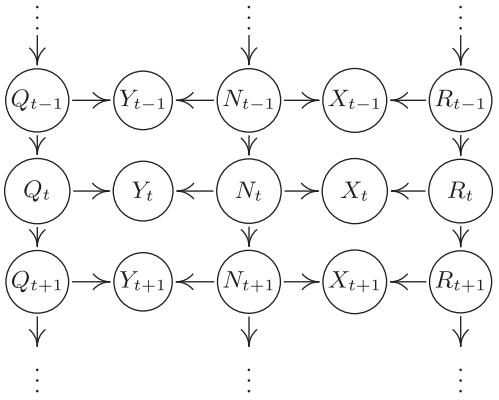

Fig. 3. Special case of the time series model. $(Q_t)_t$, $(N_t)_t$, and $(R_t)_t$ are (jointly) independent of each other, but each one may be autocorrelated. Regressing Y_t on $(X_t)_t$ unblocks only paths avoiding Q_t: the estimate $\hat{Y}_t := Y_t - E[Y_t|(X_t)_t]$ of Y_t does not regress out any variability caused by Q_t and is more accurate than the i.i.d. estimate $Y_t - E[Y_t|X_t]$. Regressing Y_t on Y_{t-1} unblocks two paths: one avoiding Q_t, and the other not. The first may enhance the estimate of Q_t; the second, however, may worsen it. In general, the contribution of both parts cannot be separated. However, they can for particular time series $(Q_t)_t$.

Method. In the i.i.d. case, we proposed to predict each Y_t from the other stars X_t and then use the residuals as a reconstruction for Q_t (up to its mean). If we are really interested in the detection of transits $(F_t)_t$ rather than in reconstructing $Q_t = h(S_t, F_t)$, we can attempt to filter out the autoregressive (AR) component of $(Y_t)_t$ that comes from $(f(N_t))_t$ and $(S_t)_t$, as long as this does not affect F_t. Consider $\alpha, \delta \in \mathbb{N}_{>0}$ with $\mathbb{P}(\Delta < \alpha) = 1$, as above. Here, δ specifies the size of windows in past and future that we will use as regression inputs. Define $\mathcal{W} := [-\alpha - \delta, -\alpha] \cup [\alpha, \alpha + \delta]$. We further write $Y_{t+\mathcal{W}} := (Y_{t+s})_{s\in\mathcal{W}}$. The method consists of the following steps.

Because the signal of interest F_t differs from c only on a compact support, we are able to prove that the method does not destroy any relevant information about the transits in F_t, as long as we carefully choose the parameters α and δ.

Half-Sibling Regression for Time Series

i) Choose a test set containing those points where we want to predict, with indices $\mathcal{I}_{\text{test}}$. Construct a training set, with indices $\mathcal{I}_{\text{train}}$, containing all those points that are separated by more than $\alpha + \delta$ from the test set.

ii) Regress Y_t on $Y_{t+\mathcal{W}}$ and X_t using all $t \in \mathcal{I}_{\text{train}}$ for training and hyperparameter tuning. Use the resulting model for predicting $\hat{Y}_t$ from $Y_{t+\mathcal{W}}$ and X_t for all $t \in \mathcal{I}_{\text{test}}$.

In principle, $\mathcal{I}_{\text{test}}$ may be a singleton, in which case we build a model for a single test point. If we have to do this for every possible choice of that point, computational requirements become rather large. In practice, we thus use a test set that contains roughly a third of the data, which means we need to carry out the above procedure three times to build models for all possible test points.

Proposition 6. *Assume that, for any $t \in \mathbb{Z}$, we have $Y_t = f(N_t) + h(S_t, F_t)$ with F_t as in* [9]. *Assume further that $\hat{Y}_t := \phi(Y_{t+\mathcal{W}}, X_t)$ with $\phi(y,x) = E[Y_t|Y_{t+\mathcal{W}} = y, X_t = x]$ constructed by half-sibling regression for time series as described above*; *here, ϕ is well defined if we assume that the conditional distribution of Y_t given $f(N_{t+\mathcal{W}}) + S_{t+\mathcal{W}}$ and X_t does not depend on t. Then*

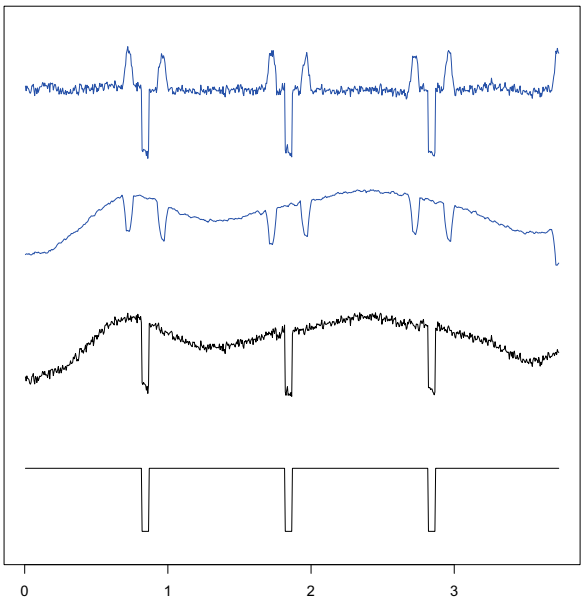

Fig. 4. Simulated transit reconstruction using half-sibling regression for time series, but without regressing on X_t. From bottom to top: $(F_t)_t$, $(Y_t)_t$, $(\hat{Y}_t)_t$ with $Y_t = S_t F_t$, and $\hat{Q}_t := (Y_t - \hat{Y}_t)_t$. The estimate $\hat{Y}_t$ was trained using ridge regression with regularization parameter $\lambda = 0.1$. Transits were also present in the training set. Note that the transit itself is preserved. However, some artifacts (here: bumps) are introduced to the right and left of the transit.

$$\hat{Y}_t \perp\!\!\!\perp F_t \quad \text{for all } t \in \mathcal{I}_{\text{transit}}. \qquad [10]$$

As a consequence, we can use $\hat{Y}_t$ to correct the observed Y_t, and we never remove any information about the transit F_t. The proof is immediate because, for a fixed $t_{test} \in \mathcal{I}_{\text{transit}}$, we have that Y_t, $Y_{t+\mathcal{W}}$, and X_t for $t \in \mathcal{I}_{\text{train}}$, as well as $Y_{t_{test}+\mathcal{W}}$ and $X_{t_{test}}$, are independent of $(\Delta_\tau(F_t)_{t \in \mathcal{I}_{\text{transit}}})$.

In general, $\hat{Y}_t$ will not be independent of F_t for $t \notin \mathcal{I}_{\text{transit}}$; in other words, correction of Y_t using $\hat{Y}_t$ (e.g., by subtraction) may distort the signal outside of the transit time window $\mathcal{I}_{\text{transit}}$. This is visualized in Fig. 4. In practice, the training set usually contains more than one transit. In this case, we cannot prove [10]. However, we would only expect a real distortion of the transit signal if the transits are abundant and thus form a significant fraction of the training set; again, see Fig. 4.

Applications

Synthetic I.I.D. Data. We first analyze two simulated data sets.

***Increasing relative strength of* N *in a single* X.** We consider 20 instances (each time we sample 200 i.i.d. data points) of the model $Y = f(N) + Q$ and $X = g(N) + R$, where f and g are randomly chosen sigmoid functions and the variables N, Q, and R are normally distributed. The SD for R is chosen uniformly between 0.05 and 1, and the SD for N is between 0.5 and 1. Because Q can be recovered only up to a shift in the mean, we set its sample mean to zero. The distribution for R, however, has a mean that is chosen uniformly between -1 and 1, and its SD is chosen from the vector $(1, 0.5, 0.25, 0.125, 0.0625, 0.03125, 0)$. Proposition 4 shows that, with decreasing SD of R, we can recover the signal Q. SD zero corresponds to the case of complete information (Proposition 2). For regressing Y on X, we use the function gam (penalized regression splines) from the R package mgcv; Fig. 5 shows that this asymptotic behavior can be seen on finite data sets.

Increasing number of observed* X_i *variables. Here, we consider the same simulation setting as before, this time simulating $X_i = g_i(N) + R_i$ for $i = 1, \ldots, d$. We have shown, in Proposition 5, that, if the number of variables X_i tends to infinity, we are able to reconstruct the signal Q. In this experiment, the SD for R_i and Q is chosen uniformly between 0.05 and 1; the distribution of N is the same as above. It is interesting to note that even additive models (in the predictor variables) work as a regression method (we use the function gam from the R package mgcv on all variables $X_1, \ldots, X_d$ and its sum $X_1 + \ldots + X_d$). Fig. 5 shows that, with increasing d, the reconstruction of Q improves.

Exoplanet Light Curves. The field of exoplanet search has recently become one of the most popular areas of astronomy research. This is largely due to the Kepler space observatory launched in 2009. Kepler observed a small fraction of the Milky Way in search of exoplanets. The telescope was pointed at the same patch of sky for more than 4 y (Fig. 6 and Fig. S2). In that patch, it monitored the brightness of 150,000 stars (selected from among 3.5 million stars in the search field), taking a stream of half-hour exposures using a set of Charge-Coupled Device (CCD) imaging chips.

Kepler detects exoplanets using the transit method. Whenever a planet passes in front of their host star(s), we observe a tiny dip in the light curve (Fig. S3). This signal is rather faint, and, for our own planet as seen from space, it would amount to a brightness change smaller than 10^{-4}, lasting less than half a day, taking place once a year, and visible from about half a percent of all directions. The level of required photometric precision to detect such transits is one of the main motivations for performing these observations in space, where they are not disturbed by atmospheric effects and it is possible to observe the same patch almost continuously using the same instrument.

For planets orbiting stars in the habitable zone (allowing for liquid water) of stars similar to the Sun, we would expect the signal to be observable, at most, every few months. We thus have very few observations of each transit. However, it has become clear that there are a number of confounders introduced by spacecraft and telescope leading to systematic changes in the light curves that are of the same magnitude or larger than the required accuracy. The dominant error is pointing jitter: If the camera field moves by a fraction of a pixel (for Kepler, the order of magnitude is 0.01 pixels), then the light distribution on the pixels will change. Each star affects a set of pixels, and we integrate their measurements to get an estimate of the star's overall brightness. Unfortunately, the pixel sensitivities are not precisely identical, and, even though one can try to correct for this, we are left with significant systematic errors. Overall, although Kepler is highly optimized for stable photometric measurements, its accuracy falls short of what is

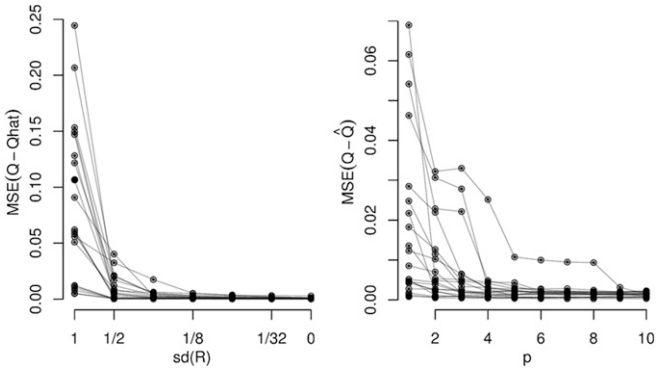

Fig. 5. (*Left*) We observe a variable $X = g(N) + R$ with invertible function g. If the variance of R decreases, the reconstruction of Q improves because it becomes easier to remove the influence $f(N)$ of the noise N from the variable $Y = f(N) + Q$ by using X; see Proposition 4. (*Right*) A similar behavior occurs with increasing the number d of predictor variables $X_i = g_i(N) + R_i$; see Proposition 5. Both plots show 20 scenarios, each connected by a thin line.

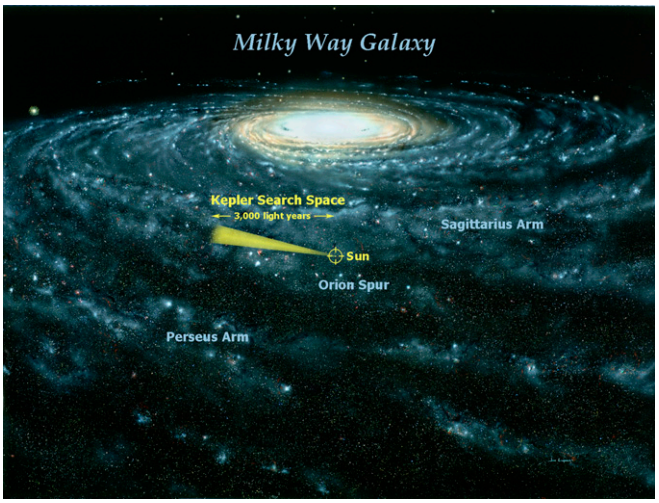

Fig. 6. View of the Milky Way with position of the Sun and depiction of the Kepler search field. Image courtesy of © Jon Lomberg.

required for reliably detecting Earth-like planets in habitable zones of Sun-like stars.

We obtained the data from the Mikulski Archive for Space Telescopes (see archive.stsci.edu/index.html). Our system, which we abbreviate as cpm (Causal Pixel Model), is based on the assumption that stars on the same CCD share systematic errors. If we pick two stars on the same CCD that are far away from each other, they will be light years apart in space, and no physical interaction between them can take place. The light curves (Fig. S4) nevertheless have similar trends, caused by systematics. We use linear ridge regression (see *Supporting Information*) to predict the light curve of each pixel belonging to the target star as a linear combination of a set of predictor pixels. Specifically, we use 4,000 predictor pixels from about 150 stars, which are selected to be closest in magnitude to the target star. (The exact number of stars varies with brightness, as brighter stars have larger images on the CCD and thus more pixels.) This is done because the systematic effects of the instruments depend somewhat on the star brightness; e.g., when a star saturates a pixel, blooming takes place, and the signal leaks to neighboring pixels. To rule out any direct optical cross-talk by stray light, we require that the predictor pixels are from stars sufficiently far away from the target star (at least 20 pixels distance on the CCD), but we always take them from the same CCD (note that Kepler has a number of CCDs, and we expect that systematic errors depend on the CCD). We train the model separately for each month, which contains about 1,300 data points. (The data come in batches, which are separated by larger errors, since the spacecraft needs to periodically redirect its antenna to send the data back to Earth.) Standard ℓ_2 regularization is used to avoid overfitting, and parameters (regularization parameter and number of input pixels) were optimized using cross-validation. Nonlinear kernel regression was also evaluated but did not lead to better results. This may be due to the fact that the set of predictor pixels is relatively large (compared with the training set size), and, among this large set, it seems that there are sufficiently many pixels that are affected by the systematics in a rather similar way as the target.

If we treat the data as i.i.d. data, our method removes some of the intrinsic variability of the target star. This is due to the fact that the signals are not i.i.d., and time acts as a confounder. If, among the predictor stars, there exists one whose intrinsic variability is very similar to the target star, then the regression can attenuate variability in the latter. This is unlikely to work exactly, but, given the limited observation window, an approximate match (e.g., stars varying at slightly different frequencies) will already lead to some amount of attenuation. Because exoplanet transits are rare, it is very unlikely (but not impossible) that the same mechanism will remove some transits.

Note that, for the purpose of exoplanet search, the stellar variability can be considered a confounder as well, independent of the planet positions that are causal for transits. To remove this, we use as additional regression inputs also the past and future of the target star (see Proposition 6). This adds an AR component to our model, removing more of the stellar variability and thus increasing the sensitivity for transits. In this case, we select an exclusion window around the point of time being corrected, to ensure that we do not remove the transit itself. Below, we report results where the AR component uses as inputs the three closest future and the three closest past time points, subject to the constraint that a window of ±9 h around the considered time point is excluded. Choosing this window corresponds to the assumption that time points earlier than −9 h or later than +9 h are not informative for the transit itself. Smaller windows allow more accurate prediction, at the risk of damaging slow transit signals. Our code is available at https://github.com/jvc2688/KeplerPixelModel.

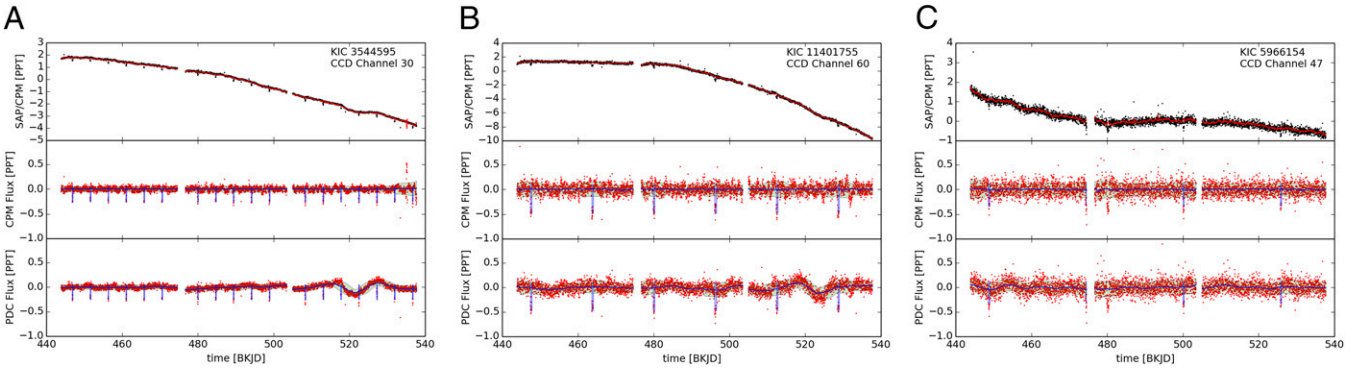

Fig. 7. Corrected fluxes using our method, for three example stars, spanning the main magnitude (brightness) range encountered. We consider a bright star (*A*), a star of moderate brightness (*B*), and a relatively faint star (*C*). SAP stands for Simple Aperture Photometry (in our case, a relative flux measure computed from summing over the pixels belonging to a star). In *A*–*C*, *Top* shows the SAP flux (black) and the cpm regression (red), i.e., our prediction of the star from other stars. *Middle* shows the cpm flux corrected using the regression (for details, see *Applications, Exoplanet Light Curves*), and *Bottom* shows the PDC flux (i.e., the default method). The cpm flux curve preserves the exoplanet transits (little downward spikes), while removing a substantial part of the variability present in the PDC flux. All *x* axes show time, measured in days since January 1, 2009.

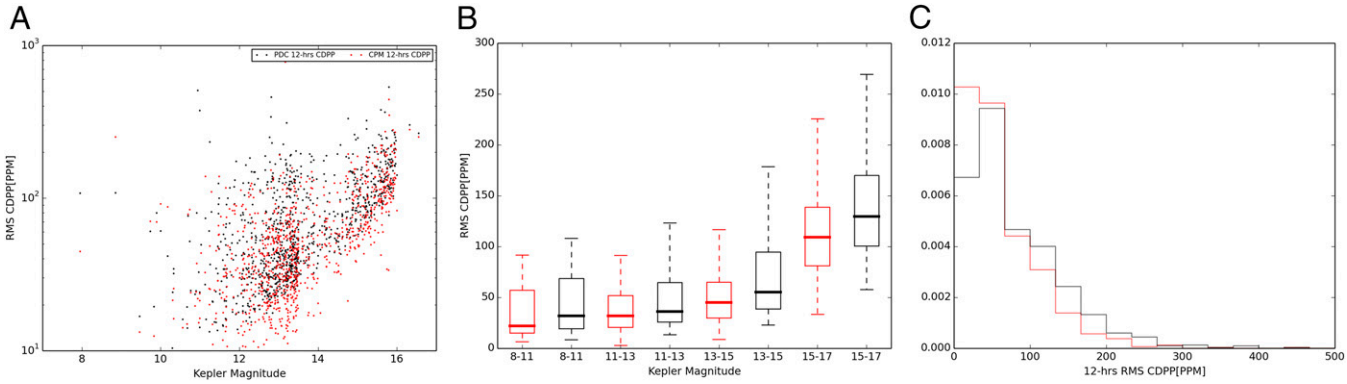

Fig. 8. Comparison of the proposed method (cpm) to the Kepler PDC method in terms of CDPP (see *Applications, Exoplanet Light Curves*). *A* shows our performance (red) vs. the PDC performance in a scatter plot, as a function of star magnitude (note that larger magnitude means fainter stars, and smaller values of CDPP indicate a higher quality as measured by CDPP). *B* bins the same dataset and shows box plots within each bin, indicating median, top quartile, and bottom quartile. The red box corresponds to cpm, and the black box refers to PDC. *C* shows a histogram of CDPP values. Note that the red histogram has more mass toward the left (i.e., smaller values of CDPP), indicating that our method overall outperforms PDC, the Kepler "gold standard."

To give a view on how our method performs, cpm is applied on several stars with known transit signals. After that, we compare them with the Kepler Presearch Data Conditioning (PDC) method (see keplergo.arc.nasa.gov/PipelinePDC.shtml). PDC builds on the idea that systematic errors have a temporal structure that can be extracted from ancillary quantities. The first version of PDC removed systematic errors based on correlations with a set of ancillary engineering data, including temperatures at the detector electronics below the CCD array, and polynomials describing centroid motions of stars. The current PDC (18, 19) performs principal component analysis (PCA) on filtered light curves of stars, projects the light curve of the target star on a PCA subspace, and subsequently removes this projection. PCA is performed on a set of relatively quiet stars close in position and magnitude. For non-i.i.d. data, this procedure could remove temporal structure of interest. To prevent this, the PCA subspace is restricted to eight dimensions, strongly limiting the capacity of the model (20).

In Fig. 7, we present corrected light curves for three typical stars of different magnitudes, using both cpm and PDC. In our theoretical analysis for the i.i.d. case, we dealt with additive noise, and could deal with multiplicative noise, e.g., by log transforming. In practice, neither of the two models is correct for our application. If we are interested in the transit (and not the stellar variability), then the variability is a multiplicative confounder. At the same time, other noises may be better modeled as additive (e.g., CCD noise). In practice, we calibrate the data by dividing by the regression estimate and then subtracting 1, i.e.,

$$\frac{Y}{E[Y|x]} - 1 = \frac{Y}{E[Y|x]} - \frac{E[Y|x]}{E[Y|x]} = \frac{Y - E[Y|x]}{E[Y|x]}.$$

Effectively, we thus perform a subtractive normalization, followed by a divisive one. This works well, taking care of both types of contaminations.

The results illustrate that our approach removes a major part of the variability present in the PDC light curves, while preserving the transit signals. To provide a quantitative comparison, we ran cpm on 1,000 stars from the whole Kepler input catalog (500 chosen randomly from the whole list, and 500 random G-type Sun-like stars), and estimated the Combined Differential Photometric Precision (CDPP) for cpm and PDC. CDPP is an estimate of the relative precision in a time window, indicating the noise level seen by a transit signal with a given duration. The duration is typically chosen to be 3 h, 6 h, or 12 h (21). Shorter durations are appropriate for planets close to their host stars, which are the ones that are easier to detect. We use the 12-h CDPP metric, because the transit duration of an Earth-like planet is roughly 10 h. Fig. 8 presents our CDPP comparison of cpm and PDC, showing that our method outperforms PDC. This is no small feat, because PDC incorporates substantial astronomical knowledge (e.g., in removing systematic trends). It should be noted, however, that PDC's runtime is much smaller than that of cpm.

Conclusion

We have assayed half-sibling regression, a simple yet effective method for removing the effect of systematic noise from observations. It uses the information contained in a set of other observations affected by the same noise source. We have analyzed both i.i.d. and time series data. The main motivation for the method was its application to exoplanet data processing, which we discussed in some detail, with rather promising results. However, we expect that it will have applications in other domains as well.

Our method may enable astronomical discoveries at higher sensitivity on the existing Kepler satellite data. Moreover, we anticipate that methods to remove systematic errors will further increase in importance: By May 2013, two of the four reaction wheels used to control the Kepler spacecraft were dysfunctional, and, in May 2014, NASA announced the K2 mission, using the remaining two wheels in combination with thrusters to control the spacecraft and continue the search for exoplanets in other star fields. Systematic errors in K2 data are significantly larger because the spacecraft has become harder to control. In addition, NASA is planning the launch of another space telescope for 2017. The Transiting Exoplanet Survey Satellite (tess.gsfc.nasa.gov/) will perform an all-sky survey for small (Earth-like) planets of nearby stars. To date, no Earth-like planets orbiting Sun-like stars in the habitable zone have been found. This is likely to change in the years to come, which would be a major scientific discovery. In particular, although the proposed method treats the problem of removing systematic errors as a preprocessing step, we are also exploring the possibility of jointly modeling systematics and transit events. This incorporates additional knowledge about the events that we are looking for in our specific application, and it has already led to promising results (20).

ACKNOWLEDGMENTS. We thank Stefan Harmeling, James McMurray, Oliver Stegle, and Kun Zhang for helpful discussion, and the anonymous reviewers for helpful suggestions and references. D.W.H., D.W., and D.F.-M. were partially supported by NSF (IIS-1124794) and NASA (NNX12AI50G) and the Moore-Sloan Data Science Environment at NYU. C.-J.S.-G. was supported by a Google Europe Doctoral Fellowship in Causal Inference. J.P. was supported by the Max Planck ETH Center for Learning Systems.

1. Schölkopf B, et al. (2012) On causal and anticausal learning. *Proceedings of the 29th International Conference on Machine Learning (ICML)*, eds Langford J, Pineau J (Omnipress, New York), pp 1255–1262.
2. Pearl J (2000) *Causality* (Cambridge Univ Press, New York).
3. Spirtes P, Glymour C, Scheines R (1993) *Causation, Prediction, and Search* (MIT Press, Cambridge, MA), 2nd Ed.
4. Aldrich J (1989) Autonomy. *Oxf Econ Pap* 41(1):15–34.
5. Hoover KD (2008) Causality in economics and econometrics. *Economics and Philosophy*, eds Durlauf SN, Blume LE (Palgrave Macmillan, New York), 2nd Ed, Vol 6.
6. Peters J, Mooij J, Janzing D, Schölkopf B (2014) Causal discovery with continuous additive noise models. *J Mach Learn Res* 15(Jun):2009–2053.
7. Hoyer P, Janzing D, Mooij JM, Peters J, Schölkopf B (2009) Nonlinear causal discovery with additive noise models. *Advances in Neural Information Processing Systems*, eds Koller D, Schuurmans D, Bengio Y, Bottou L (MIT Press, Cambridge, MA), Vol 21, pp 689–696.
8. Schölkopf B, et al. (2015) Removing systematic errors for exoplanet search via latent causes. *Proceedings of the 32nd International Conference on Machine Learning.*, eds Bach F, Blei D (Microtome, Brookline, MA), pp 2218–2226.
9. Taylor JR (1997) *An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements* (Univ Science Books, Herndon, VA), 2nd Ed.
10. Price AL, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 38(8):904–909.
11. Yu J, et al. (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet* 38(2):203–208.
12. Johnson WE, Li C, Rabinovic A (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 8(1):118–127.
13. Kang HM, Ye C, Eskin E (2008) Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots. *Genetics* 180(4):1909–1925.
14. Stegle O, Kannan A, Durbin R, Winn JM (2008) Accounting for non-genetic factors improves the power of eQTL studies. *Research in Computational Molecular Biology: 12th Annual International Conference, RECOMB 2008*, eds Vingron M, Wong L (Springer, New York), pp 411–422.
15. Gagnon-Bartsch JA, Speed TP (2012) Using control genes to correct for unwanted variation in microarray data. *Biostatistics* 13(3):539–552.
16. Janzing D, Peters J, Mooij J, Schölkopf B (2009) Identifying confounders using additive noise models. *25th Conference on Uncertainty in Artificial Intelligence*, eds Bilmes J, Ng AY (AUAI Press, Corvallis, OR), pp 249–257.
17. Padmanabhan N, et al. (2008) An improved photometric calibration of the Sloan Digital Sky Survey imaging data. *Astrophys J* 674(2):1217–1233.
18. Stumpe MC, et al. (2012) Kepler Presearch Data Conditioning I—Architecture and algorithms for error correction in Kepler light curves. *Publ Astron Soc Pac* 124:985–999.
19. Smith JC, et al. (2012) Kepler Presearch Data Conditioning II—A Bayesian approach to systematic error correction. *Publ Astron Soc Pac* 124:1000–1014.
20. Foreman-Mackey D, et al. (2015) A systematic search for transiting planets in the K2 data. *Astrophys J* 806(2):215.
21. Christiansen JL, et al. (2012) The derivation, properties, and value of Kepler's combined differential photometric precision. *Publ Astron Soc Pac* 124:1279–1287.
22. Jacob L, Gagnon-Bartsch JA, Speed TP (2016) Correcting gene expression data when neither the unwanted variation nor the factor of interest are observed. *Biostatistics* 17(1):16–28.
23. Hastie T, Tibshirani R, Friedman J (2009) *The Elements of Statistical Learning; Data Mining, Inference and Prediction* (Springer, New York), 2nd Ed.