Fitting a Trend Line to Noisy, Uncertain Data

A Bayesian Errors-in-Variables Isotonic Regression (BEVIR), With Application To Analysis of Sea-Level Index Points

Christopher G. Piecuch

Department of Physical Oceanography
Woods Hole Oceanographic Institution

2022

A bevir

(image from `https://en.wikipedia.org/wiki/Beaver`)

The Situation ...

... in a textbook

Suppose you have an $n \times 1$ vector of a response process $\boldsymbol{y} = [y_1, y_2, \ldots, y_n]^\mathsf{T}$ and an $n \times p$ matrix X of p predictor processes $\boldsymbol{x}_k = [x_{k,1}, y_{k,2}, \ldots, y_{k,n}]^\mathsf{T}$ for $k \in [1, p]$. And say you want to fit the linear model

$$\boldsymbol{y} = \mathsf{X}\boldsymbol{b} + \boldsymbol{e}, \tag{1}$$

where $\boldsymbol{b}$ is a $p \times 1$ vector of coefficients and $\boldsymbol{e}$ is the $n \times 1$ vector of residuals. The familiar ordinary least squares estimate of the parameter vector is

$$\hat{\boldsymbol{b}}_{\mathrm{OLS}} = \left(\mathsf{X}^\mathsf{T}\mathsf{X}\right)^{-1} \mathsf{X}^\mathsf{T}\boldsymbol{y}. \tag{2}$$

An important (but often unacknowledged) assumption here is that $\boldsymbol{y}$ and X are perfectly known (without error).

The Situation ...

... in real life

However, often we don't have access to the $\boldsymbol{y}$ and X processes. More typically we have noisy, biased, gappy, and otherwise uncertain and imperfect observations of the latent processes.

For example, consider sea-level index points, which identify the relative position of sea level in space and time based on natural archives such as salt-marsh sediments (Engelhart et al., 2011). In this case, both the response process (sea level) and the predictor process (age) feature uncertainties, for example, related to indicative meaning, sediment consolidation, and radiocarbon dating. The next slide shows an example.

How do we fit a model to such messy data? How do we, for example, estimate (with uncertainty) the rate of sea-level rise? Bayes' theorem provides a framework for tackling the problem.

The Situation

An example of messy data

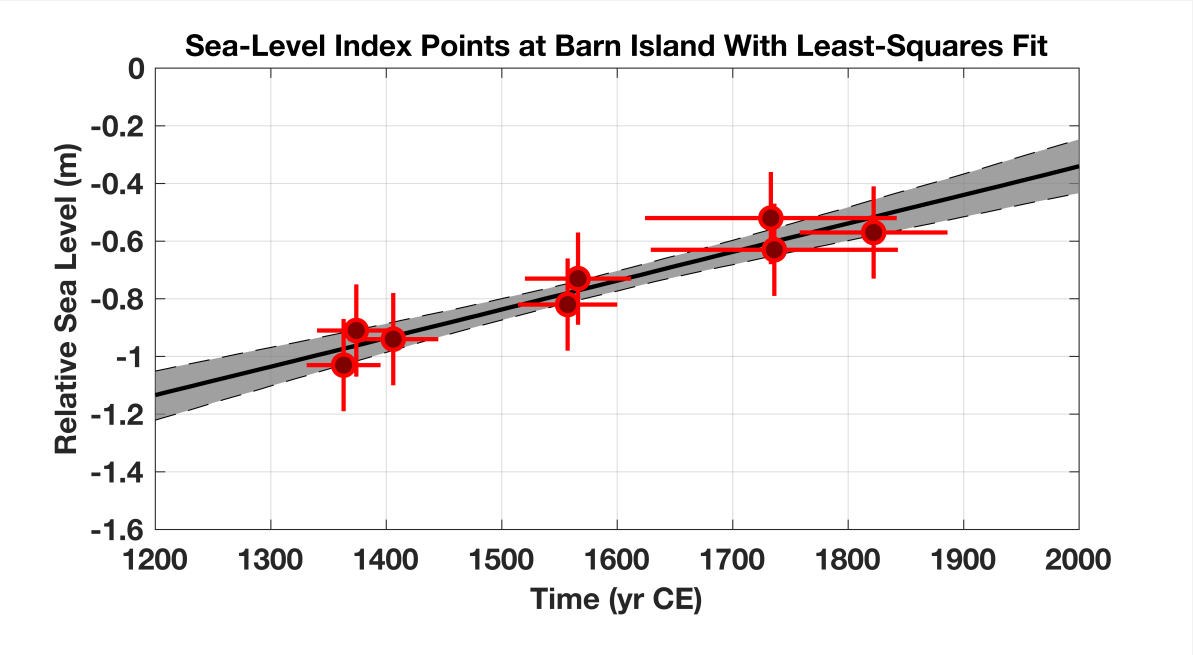

Figure 1. Red is sea-level index points. Red lines identify best estimates $\pm$ two standard errors on the sea level and age observations. Gray is the 90% confidence interval from an ordinary least squares trend fit to the data best estimates.

A Bayesian approach

Errors-In-Variables Isotonic Linear Regression

We build a Bayesian hierarchical model involving three "levels": a data level, a process level, and a prior level.

Data level

- Say we have an observation vector $\boldsymbol{z} = [z_1, z_2, \ldots, z_n]^{\mathsf{T}}$ that relates to the response process $\boldsymbol{y}$ according to

$$z_k \sim N(y_k, \delta_k^2) \ , \ \ k \in [1, n], \tag{3}$$

  where δ_k^2 is the known data error and $N(a, b)$ is the Normal distribution with mean a and variance b.
- We also have another data vector $\boldsymbol{w} = [w_1, w_2, \ldots, w_n]^{\mathsf{T}}$ that corresponds to the predictor process $\boldsymbol{x}$ following

$$w_k \sim N(x_k, \epsilon_k^2), \tag{4}$$

  where ϵ_k^2 is the corresponding known observational error. (It's the fact that we represent the data of the predictor process as uncertain that makes this "errors-in-variables.")

A Bayesian approach

Errors-In-Variables Isotonic Linear Regression

Process level

- In analogy to Equation (1), we assume that the response process ***y*** is a linear function of the predictor process ***x***

$$y_k \sim N(\alpha x_k + \beta, \gamma^2) \tag{5}$$

where α is the slope, β is the intercept, and γ^2 is the variance of the residuals, all of which are unknown.

- We also assume the monotonic ordering of the predictor

$$x_k \sim U(x_{k-1}, \infty), \tag{6}$$

where $U(a, b)$ is the Uniform distribution with upper and lower bounds a and b, respectively. (It's the constraint of monotonic ordering that makes this approach "isotonic.")

A Bayesian approach

Errors-In-Variables Isotonic Linear Regression

Prior level

- To close the model, we impose priors on the parameters

$$\alpha \sim N(\tilde{\mu}, \tilde{\zeta}^2), \tag{7}$$
$$\beta \sim N(\tilde{\eta}, \tilde{\sigma}^2), \tag{8}$$
$$\gamma^2 \sim G^{-1}(\tilde{\xi}, \tilde{\chi}), \tag{9}$$

where $G^{-1}(a, b)$ is Inverse-Gamma distribution with shape parameter a and inverse scale b, and tildes distinguish the (fixed) hyperparameters from the (uncertain) parameters.

Posterior distribution

- From Bayes' rule and the model equations, we assume the posterior of the process and parameters given the data is

$$p(\boldsymbol{y}, \boldsymbol{x}, \alpha, \beta, \gamma^2 \big| \boldsymbol{z}, \boldsymbol{w}) \propto p(\alpha)p(\beta)p(\gamma^2)$$
$$\times \prod_{k=1}^{n} \left[p(z_k \big| y_k) p(w_k \big| x_k) p(x_k \big| x_{k-1}) p(y_k \big| x_k, \alpha, \beta, \gamma) \right] \tag{10}$$

A Bayesian approach

Errors-In-Variables Isotonic Linear Regression

Generating solutions

- To evaluate the model, we use a Markov chain Monte Carlo (Gelman et al., 2013; Wikle and Berliner, 2007).
- We use a Gibbs sampler, iteratively drawing from the full conditional distributions of each process and parameter.
- Full conditionals are determined by considering only terms in the posterior distribution (Equation 10) that include the process or parameter of interest (all else being constant).
- For example, symbolically, the full conditional for α is
  1. $p(\alpha|\boldsymbol{y}, \boldsymbol{x}, \beta, \gamma^2 \boldsymbol{z}, \boldsymbol{w}) = p(\alpha) \prod_{k=1}^{n} \left[p(y_k|x_k, \alpha, \beta, \gamma) \right]$
- Full conditionals are written out explicitly in the Appendix.

A Bayesian approach

Errors-In-Variables Isotonic Linear Regression

Generating solutions

- Following Wikle and Berliner (2007), the Gibbs sampler can be prescribed algorithmically as follows:
  1. Initialize $\boldsymbol{y}^{(0)}, \boldsymbol{x}^{(0)}, \alpha^{(0)}, \beta^{(0)}, \gamma^{2,(0)}$.
  2. Iterate by generating $\boldsymbol{y}^{(i+1)}, \boldsymbol{x}^{(i+1)}, \alpha^{(i+1)}, \beta^{(i+1)}, \gamma^{2,(i+1)}$ given $\boldsymbol{y}^{(i)}, \boldsymbol{x}^{(i)}, \alpha^{(i)}, \beta^{(i)}, \gamma^{2,(i)}$ by sampling according to:
     - $p(\boldsymbol{y}^{(i+1)} \big| \boldsymbol{x}^{(i)}, \alpha^{(i)}, \beta^{(i)}, \gamma^{2,(i)}, \boldsymbol{z}, \boldsymbol{w})$
     - $p(\boldsymbol{x}^{(i+1)} \big| \boldsymbol{y}^{(i+1)}, \alpha^{(i)}, \beta^{(i)}, \gamma^{2,(i)}, \boldsymbol{z}, \boldsymbol{w})$
     - $p(\alpha^{(i+1)} \big| \boldsymbol{y}^{(i+1)}, \boldsymbol{x}^{(i+1)}, \beta^{(i)}, \gamma^{2,(i)}, \boldsymbol{z}, \boldsymbol{w})$
     - $p(\beta^{(i+1)} \big| \boldsymbol{y}^{(i+1)}, \boldsymbol{x}^{(i+1)}, \alpha^{(i+1)}, \gamma^{2,(i)}, \boldsymbol{z}, \boldsymbol{w})$
     - $p(\gamma^{2,(i+1)} \big| \boldsymbol{y}^{(i+1)}, \boldsymbol{x}^{(i+1)}, \alpha^{(i+1)}, \beta^{(i+1)}, \boldsymbol{z}, \boldsymbol{w})$

     where $i \in [1, I_{\mathrm{burn}} + I_{\mathrm{post}}]$.
- Delete the first I_{burn} "burn-in" draws (to reduce the effects of the initial transient adjustment).
- Thin the remaining I_{post} draws by only keeping every I_{thin} sample (to reduce autocorrelation of samples).

A Bayesian approach

An example: sea-level index points

- Consider the sea-level index points in Figure 1, which identify sea-level rise based on salt-marsh sediment from Barn Island, Connecticut (USA) during ∼1300–1850 CE.

$$\boldsymbol{z} = \begin{bmatrix} -1.03 \\ -0.94 \\ -0.91 \\ -0.82 \\ -0.73 \\ -0.63 \\ -0.57 \\ -0.52 \end{bmatrix}; \boldsymbol{\delta} = \begin{bmatrix} 0.16 \\ 0.16 \\ 0.16 \\ 0.16 \\ 0.16 \\ 0.16 \\ 0.16 \\ 0.16 \end{bmatrix}; \boldsymbol{w} = \begin{bmatrix} 1363 \\ 1406 \\ 1374 \\ 1557 \\ 1566 \\ 1736 \\ 1822 \\ 1733 \end{bmatrix}; \boldsymbol{\epsilon} = \begin{bmatrix} 32 \\ 39 \\ 34 \\ 43 \\ 46 \\ 107 \\ 64 \\ 109 \end{bmatrix}$$

- $\boldsymbol{z}$ and $\boldsymbol{\delta}$ are in meters while $\boldsymbol{w}$ and $\boldsymbol{\epsilon}$ are in years (CE).
- The data are taken from the Engelhart and Horton (2012) compilation (originally reported by Donnelly et al., 2004).

A Bayesian approach

An example: sea-level index points

- Using least squares to fit a trend through the best estimates of the observations, we obtain a rate of 1.0 ± 0.2 mm/yr (95% confidence interval; Figure 1).
  1. We have the "gut sense" that this formal uncertainty is too small given the large errors in the data.
- To evaluate the Bayesian model we:
  1. use $\alpha \sim N(0, 10^{-3})$, $\beta \sim N(0, 1)$, $\gamma^2 \sim G^{-1}(0.5, 0.02)$.
  2. use $I_{\text{burn}} = 1000$, $I_{\text{post}} = 10000$, and $I_{\text{thin}} = 10$.[1]
- Posterior solutions are shown in Figures 2 and 3 below.
- Matlab code used the produce the results is available at `https://github.com/christopherpiecuch/bevir`.
- The Bayesian rate (0.6 ± 0.7 mm/yr; 95% credible interval) is more uncertain and muted, but arguably more realistic given the errors in the data.

[1] This takes a few seconds to run on a 2017 MacBook Pro with a 3.1 GHz Quad-Core Intel Core i7.

A Bayesian approach

An example: sea-level index points

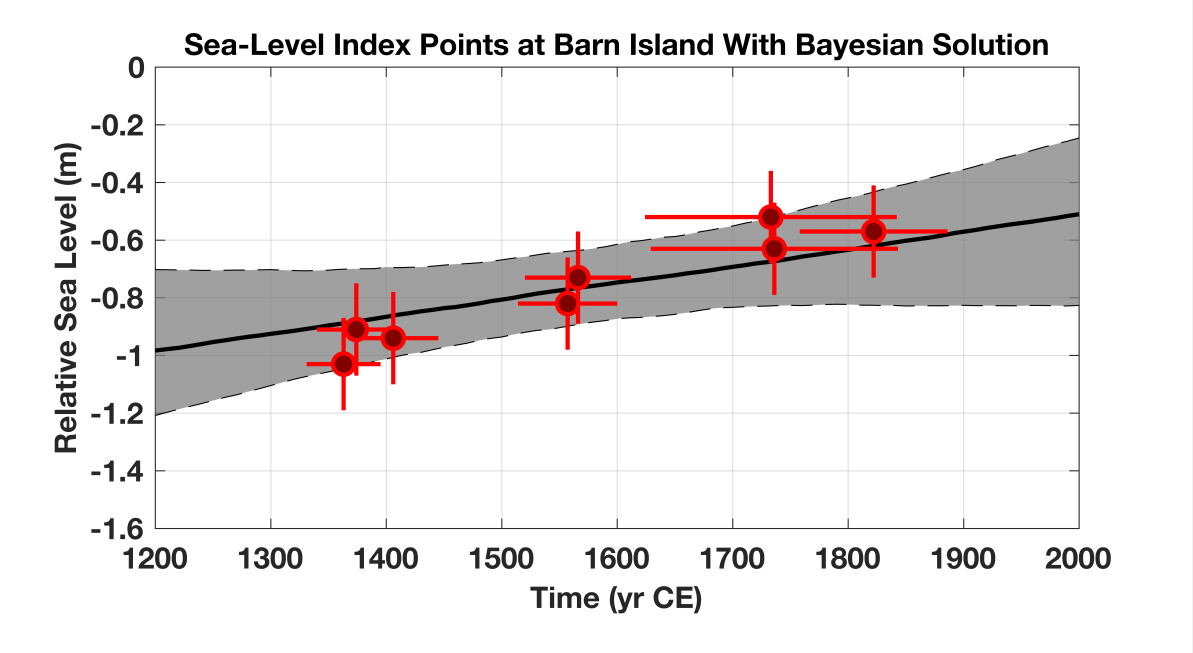

Figure 2. Red is sea-level index points. Red lines identify best estimates $\pm$ two standard errors on the sea level and age observations. Gray is the 90% posterior credible interval from the Bayesian model (solid black is the median estimate).

A Bayesian approach

An example: sea-level index points

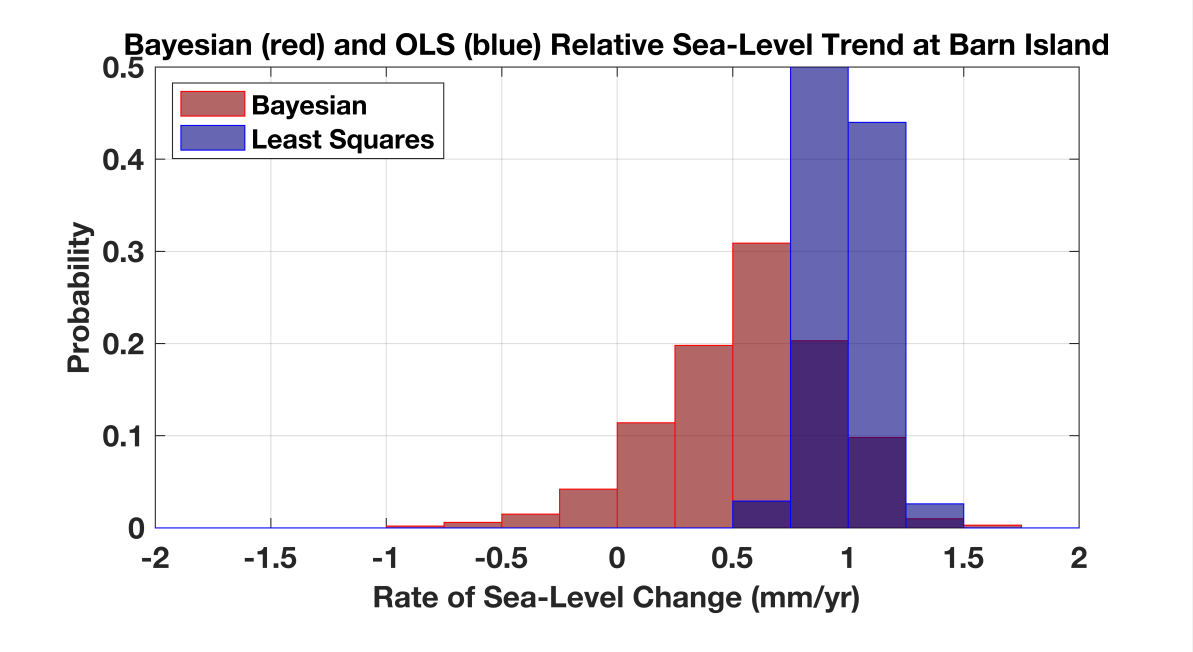

Figure 3. Histograms of Bayesian (red; 0.6 ± 0.7 mm/yr; 95% credible interval) and least squares (blue; 1.0 ± 0.2 mm/yr; 95% confidence interval) sea-level trend estimates.

Appendix

Full conditional distributions

Slope α

- $\alpha\big|\cdot \sim N(\psi_\alpha V_\alpha, \psi_\alpha)$, with,
- $V_\alpha \doteq \tilde{\zeta}^{-2}\tilde{\mu} + \gamma^{-2}\sum_{k=1}^{n} x_k(y_k - \beta)$, and,
- $\psi_\alpha \doteq (\tilde{\zeta}^{-2} + \gamma^{-2}\sum_{k=1}^{n} x_k^2)^{-1}$.

Intercept β

- $\beta\big|\cdot \sim N(\psi_\beta V_\beta, \psi_\beta)$, with,
- $V_\alpha \doteq \tilde{\sigma}^{-2}\tilde{\eta} + \gamma^{-2}\sum_{k=1}^{n}(y_k - \alpha x_k)$, and,
- $\psi_\beta \doteq (\tilde{\sigma}^{-2} + n\gamma^{-2})^{-1}$.

Variance γ^2

- $\gamma^2\big|\cdot \sim G^{-1}[\tilde{\xi} + \frac{n}{2}, \tilde{\chi} + \frac{1}{2}\sum_{k=1}^{n}(y_k - \alpha x_k - \beta)^2]$.

Appendix

Full conditional distributions (continued)

Response y_k for $k \in [1, n]$

- $y_k \big| \cdot \sim N(\psi_y V_y, \psi_y)$, with,
- $V_y \doteq \delta^{-2} z_k + \gamma^{-2}(\alpha x_k + \beta)$, and,
- $\psi_y \doteq (\delta^{-2} + \gamma^{-2})^{-1}$.

Predictor x_k for $k \in [1, n]$

- $x_k \big| \cdot \sim N_{[x_{k-1}, x_{k+1}]}(\psi_x V_x, \psi_x)$, with,
- $V_x \doteq \epsilon^{-2} w_k + \gamma^{-2}\alpha(y_k - \beta)$, and
- $\psi_x \doteq (\epsilon^{-2} + \alpha^2\gamma^{-2})^{-1}$.
- $N_{[i,j]}(a, b)$ is the truncated Normal with "mean" a, "variance" b, and upper and lower bounds i and j.
- If $k = 1$, then $x_{k-1} = -\infty$; if $k = n$, then $x_{k+1} = \infty$.

Appendix

Further reading

For more details on Bayesian models and errors-in-variables approaches in the context of sea level, see Ashe et al. (2019).

Cited references

- Ashe, E. L., et al. (2019). https://doi.org/10.1016/j.quascirev.2018.10.032
- Donnelly, J. P., et al. (2004). https://doi.org/10.1029/2003GL018933
- Engelhart, S. E., et al. (2011). https://doi.org/10.5670/oceanog.2011.28
- Engelhart, S. E., and B. P. Horton (2012). https://doi.org/10.1016/j.quascirev.2011.09.013
- Gelman, A., et al. (2013). http://www.stat.columbia.edu/ gelman/book/
- Wikle, C. K., and L. M. Berliner (2007). https://doi.org/10.1016/j.physd.2006.09.017