Bayesian regression with confounding variables

Luke Travis presenting joint work with Kolyan Ray

CFE-CMStatistics, King's College London
15th December 2024

IMPERIAL

Problem Setup

- Confounded regression model (given below for one observation):

$$y_i = f(\mathbf{x}_i) + \mathbf{h}_i^T \delta + \nu_i, \qquad i = 1, \ldots, n.$$

- $\mathbf{x}_i \in \mathbb{R}^p$ observable, $\mathbf{h}_i \in \mathbb{R}^q$ unobservable with q unknown. Both are random.

Problem Setup

- Confounded regression model (given below for one observation):

$$y_i = f(\mathbf{x}_i) + \mathbf{h}_i^T \delta + \nu_i, \qquad i = 1, \ldots, n.$$

- $\mathbf{x}_i \in \mathbb{R}^p$ observable, $\mathbf{h}_i \in \mathbb{R}^q$ unobservable with q unknown. Both are random.
- $\Gamma := \text{cov}(\mathbf{h}_i, \mathbf{x}_i) \neq 0.$
- ν_i noise.
- **Goal: to perform inference on f.**

Problem Setup

- With $X = (\mathbf{x}_1^T, \ldots, \mathbf{x}_n^T)^T$ and $H = (\mathbf{h}_1^T, \ldots, \mathbf{h}_n^T)^T$, the model for n observations is

$$Y = f(X) + H\delta + \nu, \quad \textbf{(CM)}$$

$f(X) = (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n))^T.$

Problem Setup

- With $X = (\mathbf{x}_1^T, \ldots, \mathbf{x}_n^T)^T$ and $H = (\mathbf{h}_1^T, \ldots, \mathbf{h}_n^T)^T$, the model for n observations is

$$Y = f(X) + H\delta + \nu, \quad \textbf{(CM)}$$

  $f(X) = (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n))^T$.

- W.l.o.g. $\text{Cov}(\mathbf{h}_i) = I_q$ (otherwise take $H \mapsto H\text{Cov}(H)^{-1/2}$ and $\delta \mapsto \text{Cov}(H)^{1/2}\delta$).

Problem Setup

- With $X = (\mathbf{x}_1^T, \ldots, \mathbf{x}_n^T)^T$ and $H = (\mathbf{h}_1^T, \ldots, \mathbf{h}_n^T)^T$, the model for n observations is

$$Y = f(X) + H\delta + \nu, \quad \textbf{(CM)}$$

  $f(X) = (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n))^T$.

- W.l.o.g. $\text{Cov}(\mathbf{h}_i) = I_q$ (otherwise take $H \mapsto H\text{Cov}(H)^{-1/2}$ and $\delta \mapsto \text{Cov}(H)^{1/2}\delta$).

- Recalling that $\Gamma = \text{Cov}(\mathbf{h}_i, \mathbf{x}_i) \in \mathbb{R}^{q \times p}$,

$$X = H\Gamma + E,$$

  with $\text{Cov}(\mathbf{h}_i, \mathbf{e}_i) = 0$ — (that is, the rows of each matrix are uncorrelated).

Relationship with the perturbed model

- For any $b \in \mathbb{R}^p$, the confounded model can be written (Ćevid et al., 2020)

$$Y = f(X) + Xb + \underbrace{(H\delta - Xb + \nu)}_{=:\varepsilon}$$

Ćevid, D., Bühlmann, P., Meinhausen, N. *Spectral deconfounding via perturbed sparse linear models.* JMLR, 2020.

Relationship with the perturbed model

- For any $b \in \mathbb{R}^p$, the confounded model can be written (Ćevid et al., 2020)

$$Y = f(X) + Xb + \underbrace{(H\delta - Xb + \nu)}_{=:\varepsilon}$$

- In particular, if

$$b = (\Gamma^T\Gamma + \Sigma_E)^{-1}\Gamma^T\delta,$$

then $\text{Cov}(X, \varepsilon) = 0$ so that ε **is uncorrelated with the design.**

Ćevid, D., Bühlmann, P., Meinhausen, N. *Spectral deconfounding via perturbed sparse linear models.* JMLR, 2020.

Relationship with the perturbed model

- For any $b \in \mathbb{R}^p$, the confounded model can be written (Ćevid et al., 2020)

$$Y = f(X) + Xb + \underbrace{(H\delta - Xb + \nu)}_{=:\varepsilon}$$

- In particular, if

$$b = (\Gamma^T\Gamma + \Sigma_E)^{-1}\Gamma^T\delta,$$

  then $\text{Cov}(X, \varepsilon) = 0$ so that ε **is uncorrelated with the design.**

- This gives us the perturbed regression setting

$$Y = f(X) + Xb + \varepsilon. \quad \textbf{(PM)}$$

Ćevid, D., Bühlmann, P., Meinhausen, N. *Spectral deconfounding via perturbed sparse linear models.* JMLR, 2020.

Relationship with the perturbed model

- For any $b \in \mathbb{R}^p$, the confounded model can be written (Ćevid et al., 2020)

$$Y = f(X) + Xb + \underbrace{(H\delta - Xb + \nu)}_{=:\varepsilon}$$

- In particular, if

$$b = (\Gamma^T\Gamma + \Sigma_E)^{-1}\Gamma^T\delta,$$

  then $\text{Cov}(X, \varepsilon) = 0$ so that ε **is uncorrelated with the design.**

- This gives us the perturbed regression setting

$$Y = f(X) + Xb + \varepsilon. \quad \textbf{(PM)}$$

- The **confounded model is a special case of the perturbed model**, with a relationship between b and the design.

Ćevid, D., Bühlmann, P., Meinhausen, N. *Spectral deconfounding via perturbed sparse linear models.* JMLR, 2020.

Identifiability of the perturbed model

- In general the perturbed model is not identifiable.

Identifiability of the perturbed model

- In general the perturbed model is not identifiable.
- For example if we consider the linear case

$$Y = X\beta + Xb + \varepsilon,$$

one would only be able to infer $\beta + b$.

Identifiability of the perturbed model

- In general the perturbed model is not identifiable.
- For example if we consider the linear case

$$Y = X\beta + Xb + \varepsilon,$$

  one would only be able to infer $\beta + b$.

- Therefore need to make further assumptions if working with the perturbed model: e.g. **a sparse β and a dense b** which is **small in some norm** in the case of high-dimensional regression $(n < p)$.

Identifiability of the perturbed model

- In general the perturbed model is not identifiable.
- For example if we consider the linear case

$$Y = X\beta + Xb + \varepsilon,$$

  one would only be able to infer $\beta + b$.

- Therefore need to make further assumptions if working with the perturbed model: e.g. **a sparse β and a dense b** which is **small in some norm** in the case of high-dimensional regression $(n < p)$.
- However the confounded model itself gives us something to work with because of the form

$$b = (\Gamma^T\Gamma + \Sigma_E)^{-1}\Gamma^T\delta$$

Structure and motivation for a method

- Again consider high-dimensional linear regression

$$Y = X\beta + H\delta + \nu = X\beta + Xb + \varepsilon.$$

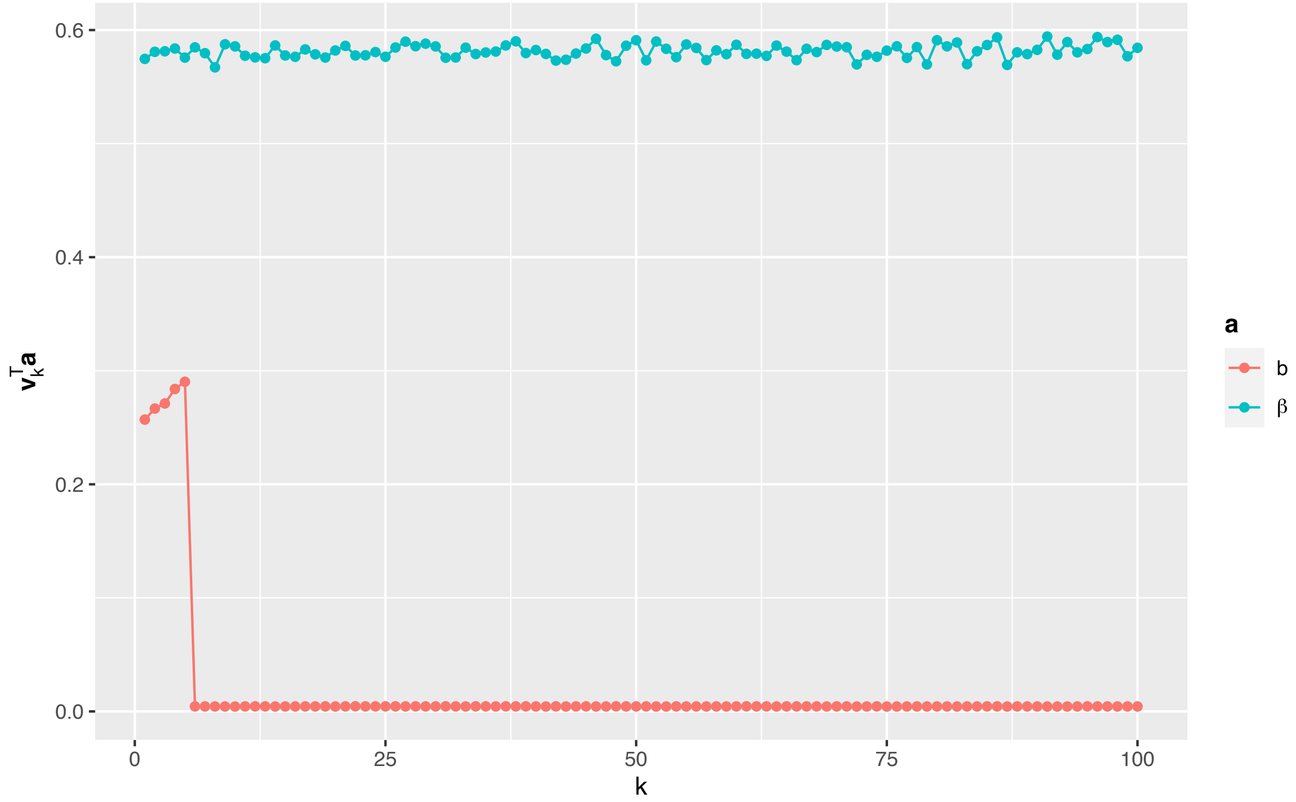
- For $X = UDV^T = \sum_{k=1}^{r} d_k \mathbf{u}_k \mathbf{v}_k^T$, and a vector $\mathbf{a}$, $\quad X\mathbf{a} = \sum_{k=1}^{r} [d_k(\mathbf{v}_k^T\mathbf{a})]\mathbf{u}_k, \qquad \|X\mathbf{a}\|_2^2 = \sum_{k=1}^{r} d_k^2(\mathbf{v}_k^T\mathbf{a})^2$

Structure and motivation for a method

- Again consider high-dimensional linear regression

$$Y = X\beta + H\delta + \nu = X\beta + Xb + \varepsilon.$$

- For $X = UDV^T = \sum_{k=1}^{r} d_k \mathbf{u}_k \mathbf{v}_k^T$, and a vector $\mathbf{a}$, $X\mathbf{a} = \sum_{k=1}^{r} [d_k(\mathbf{v}_k^T\mathbf{a})]\mathbf{u}_k$, $\|X\mathbf{a}\|_2^2 = \sum_{k=1}^{r} d_k^2(\mathbf{v}_k^T\mathbf{a})^2$



Left: Mean values of $\mathbf{v}_k^T\beta$ and $\mathbf{v}_k^T b$ (5000 realisations) from the dataset of dimension $(n, p, q) = (100,\ 200,\ 5)$ generated as

$$X = H\Gamma + E, \qquad H_{ij}, \Gamma_{ij}, E_{ij} \sim \mathcal{N}(0,1)$$

β a randomly generated sparse vector

$$\delta = (\mathcal{N}(0, \log n), \ldots, \mathcal{N}(0, \log n)) \in \mathbb{R}^5$$

$$b = (\Gamma^T\Gamma + I_p)^{-1}\Gamma^T\delta$$

IMPERIAL

Existing work (specific to sparse high-dimensional linear regression)

- In sparse high-dimensional linear regression, Ćevid et al. (2020) propose the following estimator:

$$\tilde{X} = LX, \qquad \tilde{Y} = LX$$

$$\hat{\beta} = \text{argmin}_\beta \left\{ \frac{1}{n} \|\tilde{Y} - \tilde{X}\beta\|_2^2 + \mu_1 \|\beta\|_1 \right\},$$

  for some preprocessing transform $L \in \mathbb{R}^{n \times n}$ and penalisation parameter μ_1.

- The question then is what is a good choice of L.

Ćevid, D., Bühlmann, P., Meinhausen, N. *Spectral deconfounding via perturbed sparse linear models.* JMLR, 2020.

Existing work

Some good choices of L are given by the following.

Ćevid, D., Bühlmann, P., Meinhausen, N. *Spectral deconfounding via perturbed sparse linear models.* JMLR, 2020.

Chernozhukov, V., Hansen, C., Liao, Y.. *A lava attack on the recovery of sums of dense and sparse signals.* AOS, 2017.

Existing work

Some good choices of L are given by the following.

1. The ‘trim’ transform of Ćevid et al. (2020).

   For $X = UDV^T$, let $L = U[\tilde{D}D^{-1}]V^T$, for $\tilde{d}_i = \min(\tau, d_i)$, so that $\tilde{X} = U\tilde{D}V^T$. The singular values of $\tilde{X}$ are essentially capped at some level τ.

Ćevid, D., Bühlmann, P., Meinhausen, N. *Spectral deconfounding via perturbed sparse linear models.* JMLR, 2020.

Chernozhukov, V., Hansen, C., Liao, Y.. *A lava attack on the recovery of sums of dense and sparse signals.* AOS, 2017.

Existing work

Some good choices of L are given by the following.

1. The ‘trim’ transform of Ćevid et al. (2020).

   For $X = UDV^T$, let $L = U[\tilde{D}D^{-1}]V^T$, for $\tilde{d}_i = \min(\tau, d_i)$, so that $\tilde{X} = U\tilde{D}V^T$. The singular values of $\tilde{X}$ are essentially capped at some level τ.

2. The ‘lava’ transform of Chernozukhov et al. (2017).

   Let $L = (I_n - X(X^TX + n\mu_2 I_p)^{-1}X^T)^{1/2}$. The resulting $\hat{\beta}$ is the same as the solution to the problem

$$\hat{\beta}, \hat{b} = \text{argmin}_{\beta,b} \left\{ \frac{1}{n}\|Y - X\beta - Xb\|_2^2 + \mu_1\|\beta\|_1 + \mu_2\|b\|_2^2 \right\}.$$

Ćevid, D., Bühlmann, P., Meinhausen, N. *Spectral deconfounding via perturbed sparse linear models.* JMLR, 2020.

Chernozhukov, V., Hansen, C., Liao, Y.. *A lava attack on the recovery of sums of dense and sparse signals.* AOS, 2017.

Existing work

Some good choices of L are given by the following.

1. The 'trim' transform of Ćevid et al. (2020).

   For $X = UDV^T$, let $L = U[\tilde{D}D^{-1}]V^T$, for $\tilde{d}_i = \min(\tau, d_i)$, so that $\tilde{X} = U\tilde{D}V^T$. The singular values of $\tilde{X}$ are essentially capped at some level τ.

2. The 'lava' transform of Chernozukhov et al. (2017).

   Let $L = (I_n - X(X^TX + n\mu_2 I_p)^{-1}X^T)^{1/2}$. The resulting $\hat{\beta}$ is the same as the solution to the problem

$$\hat{\beta}, \hat{b} = \text{argmin}_{\beta,b} \left\{ \frac{1}{n}\|Y - X\beta - Xb\|_2^2 + \mu_1\|\beta\|_1 + \mu_2\|b\|_2^2 \right\}.$$

   This transform also just results in an $\tilde{X}$ with transformed singular values:

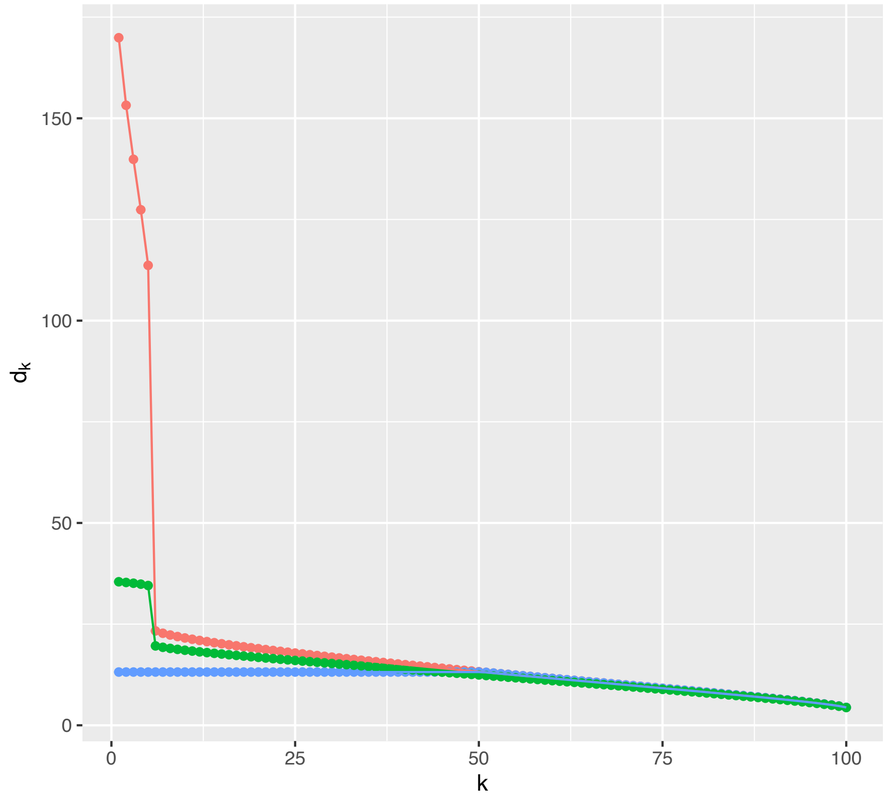
$$\tilde{d}_i = \sqrt{\frac{n\mu_2 d_i^2}{n\mu_2 + d_i^2}}$$

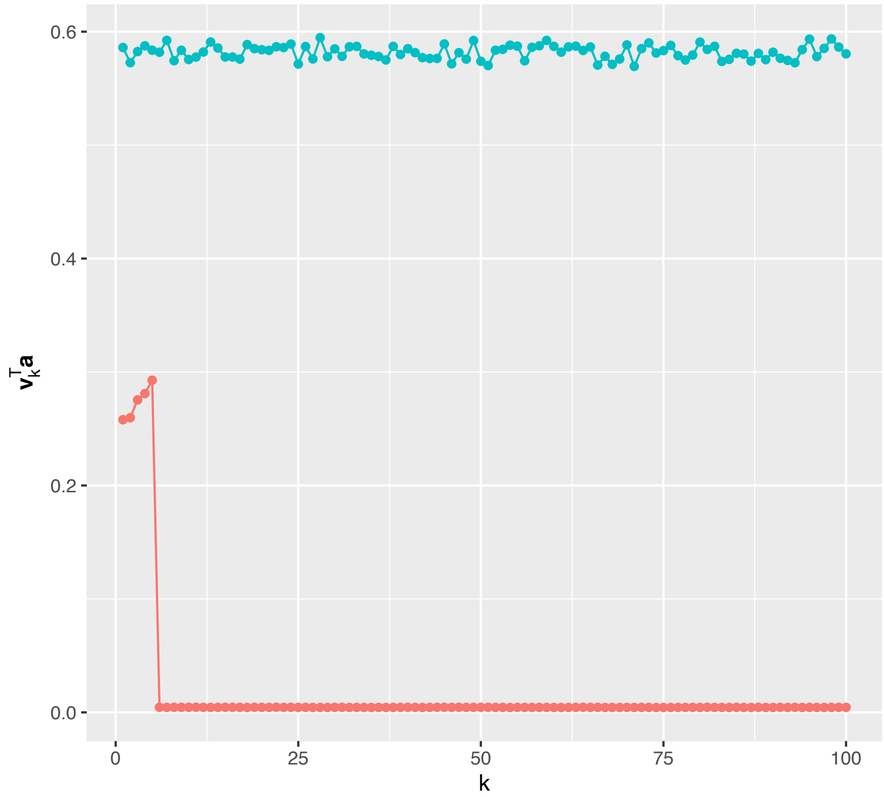
Ćevid, D., Bühlmann, P., Meinhausen, N. *Spectral deconfounding via perturbed sparse linear models.* JMLR, 2020.

Chernozhukov, V., Hansen, C., Liao, Y.. *A lava attack on the recovery of sums of dense and sparse signals.* AOS, 2017.

Example: Linear Regression

Recalling $X\mathbf{a} = \sum_{k=1}^{r} d_k \cdot (\mathbf{v}_k^T \mathbf{a})\mathbf{u}_k.$





IMPERIAL

Bayesian Approach (Likelihood and Prior)

- We consider the following Bayesian approach to both the **CM** and the **PM**.

Bayesian Approach (Likelihood and Prior)

- We consider the following Bayesian approach to both the **CM** and the **PM**.
- In both instances, our method uses the likelihood

$$\mathscr{L}(f, b \,|\, X, Y) = (2\pi\sigma_\varepsilon^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma_\varepsilon^2} \| Y - f(X) - Xb \|_2^2 \right\}$$

Bayesian Approach (Likelihood and Prior)

- We consider the following Bayesian approach to both the **CM** and the **PM**.
- In both instances, our method uses the likelihood

$$\mathscr{L}(f, b \,|\, X, Y) = (2\pi\sigma_\varepsilon^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma_\varepsilon^2} \| Y - f(X) - Xb \|_2^2 \right\}$$

- In the **PM**, b is an actual parameter, whereas in the **CM** we consider it a latent variable.

Bayesian Approach (Likelihood and Prior)

- We consider the following Bayesian approach to both the **CM** and the **PM**.
- In both instances, our method uses the likelihood

$$\mathscr{L}(f, b \,|\, X, Y) = (2\pi\sigma_\varepsilon^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma_\varepsilon^2} \| Y - f(X) - Xb \|_2^2 \right\}$$

- In the **PM**, b is an actual parameter, whereas in the **CM** we consider it a latent variable.
- For now place an arbitrary prior Π_f distribution on f.
- Throughout we place a Gaussian $\mathcal{N}_p(0, \Sigma) =: \Pi_b$ prior distribution on b.

Posterior and marginal posterior

- Marginal posterior distribution of f:

$$\Pi_f(f \in F \,|\, X, Y) = \Pi((f, b) \in F \times \mathbb{R}^p \,|\, X, Y) = \frac{\int_F \int_{\mathbb{R}^p} \mathscr{L}(f, b \,|\, X, Y) d\Pi_b(b) d\Pi_f(f)}{\int\int \mathscr{L}(f, b \,|\, X, Y) d\Pi_b(b) d\Pi_f(f)}.$$

Posterior and marginal posterior

- Marginal posterior distribution of f:

$$\Pi_f(f \in F \,|\, X, Y) = \Pi((f, b) \in F \times \mathbb{R}^p \,|\, X, Y) = \frac{\int_F \int_{\mathbb{R}^p} \mathscr{L}(f, b \,|\, X, Y) d\Pi_b(b) d\Pi_f(f)}{\int\int \mathscr{L}(f, b \,|\, X, Y) d\Pi_b(b) d\Pi_f(f)}.$$

- We can write the likelihood

$$\mathscr{L}(f, b \,|\, X, Y) = \underbrace{(2\pi\sigma_\varepsilon^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma_\varepsilon^2} \|Y - f(X)\|_2^2 \right\}}_{=:\mathscr{L}(f|X,Y)} \cdot \underbrace{\exp\left\{ -\frac{1}{2\sigma_\varepsilon^2} \|Xb\|_2^2 + \frac{1}{\sigma_\varepsilon^2} \langle Y - f(X), Xb \rangle_2 \right\}}_{=:\mathscr{L}(b|f,X,Y)}$$

Likelihood Decomposition

- Then the marginal posterior is

$$\Pi_f(F \mid X, Y) = \frac{\int_F \left[\int_{\mathbb{R}^p} \mathscr{L}(b \mid f, X, Y) d\Pi_b(b) \right] \mathscr{L}(f \mid X, Y) d\Pi_f(f)}{\int \left[\int_{\mathbb{R}^p} \mathscr{L}(b \mid f, X, Y) d\Pi_b(b) \right] \mathscr{L}(f \mid X, Y) d\Pi_f(f)}.$$

Likelihood Decomposition

- Then the marginal posterior is

$$\Pi_f(F \mid X, Y) = \frac{\int_F \left[\int_{\mathbb{R}^p} \mathscr{L}(b \mid f, X, Y) d\Pi_b(b) \right] \mathscr{L}(f \mid X, Y) d\Pi_f(f)}{\int \left[\int_{\mathbb{R}^p} \mathscr{L}(b \mid f, X, Y) d\Pi_b(b) \right] \mathscr{L}(f \mid X, Y) d\Pi_f(f)}.$$

Since $d\Pi_b(b) = \phi_p(b \mid 0, \Sigma) db$ and $\mathscr{L}(b \mid f, X, Y)$ has a Gaussian form, we can compute the terms in square brackets analytically as

$$\int_{\mathbb{R}^p} \mathscr{L}(b \mid f, X, Y) d\Pi_b(b) = \sqrt{\frac{|\Sigma_*|}{|\Sigma|}} e^{\frac{1}{2\sigma_\varepsilon^2}[Y - f(X)]^T X \Sigma_* X [Y - f(X)]},$$

with $\Sigma_* := (X^T X + \Sigma^{-1})^{-1}$.

Likelihood Decomposition

- Then the marginal posterior is

$$\Pi_f(F \mid X, Y) = \frac{\int_F \left[\int_{\mathbb{R}^p} \mathscr{L}(b \mid f, X, Y) d\Pi_b(b) \right] \mathscr{L}(f \mid X, Y) d\Pi_f(f)}{\int \left[\int_{\mathbb{R}^p} \mathscr{L}(b \mid f, X, Y) d\Pi_b(b) \right] \mathscr{L}(f \mid X, Y) d\Pi_f(f)}.$$

Since $d\Pi_b(b) = \phi_p(b \mid 0, \Sigma) db$ and $\mathscr{L}(b \mid f, X, Y)$ has a Gaussian form, we can compute the terms in square brackets analytically as

$$\int_{\mathbb{R}^p} \mathscr{L}(b \mid f, X, Y) d\Pi_b(b) = \sqrt{\frac{|\Sigma_*|}{|\Sigma|}} e^{\frac{1}{2\sigma_\varepsilon^2}[Y - f(X)]^T X \Sigma_* X [Y - f(X)]},$$

with $\Sigma_* := (X^T X + \Sigma^{-1})^{-1}$.

- Recall that $\mathscr{L}(f \mid X, Y) \propto e^{-\frac{1}{2\sigma_\varepsilon^2}[Y - f(X)]^T I_n [Y - f(X)]}$.

Likelihood Decomposition

- The marginal posterior distribution of f is

$$\Pi_f(F \mid X, Y) = \frac{\int_F e^{-\frac{1}{2\sigma_\varepsilon^2}\|L(Y-f(X))\|_2^2} d\Pi_f(f)}{\int e^{-\frac{1}{2\sigma_\varepsilon^2}\|L(Y-f(X))\|_2^2} d\Pi_f(f)},$$

  for the matrix L defined by

$$L^T L = I_n - X\Sigma_* X^T,$$

  which exists as long as Σ is symmetric positive semi-definite.

The role of Σ and L

- Writing explicitly

$$L^T L = I_n - X(X^T X + \Sigma^{-1})^{-1} X^T.$$

- Writing the SVD of $X = \sum_{k=1}^{r} \sqrt{\lambda_k}\mathbf{u}_k\mathbf{v}_k^T$, it is instructive to consider a Σ defined by

The role of Σ and L

- Writing explicitly

$$L^T L = I_n - X(X^T X + \Sigma^{-1})^{-1} X^T.$$

- Writing the SVD of $X = \sum_{k=1}^{r} \sqrt{\lambda_k}\mathbf{u}_k \mathbf{v}_k^T$, it is instructive to consider a Σ defined by

$$\Sigma = \sum_{k=1}^{p} \varphi_k \mathbf{v}_k \mathbf{v}_k^T,$$

where we have possibly extended $\{\mathbf{v}_1, \ldots, \mathbf{v}_r\}$ to an o.n. basis of $\mathbb{R}^p$ if $r < p$. In this case

$$L = \sum_{k=1}^{r} \frac{1}{\sqrt{1 + \lambda_k \varphi_k}} \mathbf{u}_k \mathbf{u}_k^T + \sum_{k=r+1}^{n} \mathbf{u}_k \mathbf{u}_k^T.$$

The role of Σ and L

- Writing explicitly

$$L^T L = I_n - X(X^T X + \Sigma^{-1})^{-1} X^T.$$

- Writing the SVD of $X = \sum_{k=1}^{r} \sqrt{\lambda_k} \mathbf{u}_k \mathbf{v}_k^T$, it is instructive to consider a Σ defined by

$$\Sigma = \sum_{k=1}^{p} \varphi_k \mathbf{v}_k \mathbf{v}_k^T,$$

where we have possibly extended $\{\mathbf{v}_1, \ldots, \mathbf{v}_r\}$ to an o.n. basis of $\mathbb{R}^p$ if $r < p$. In this case

$$L = \sum_{k=1}^{r} \frac{1}{\sqrt{1 + \lambda_k \varphi_k}} \mathbf{u}_k \mathbf{u}_k^T + \sum_{k=r+1}^{n} \mathbf{u}_k \mathbf{u}_k^T.$$

- One can see that L **shrinks vectors in the directions of span**$(\mathbf{u}_1, \ldots, \mathbf{u}_r)$ of X, and operates as the **identity on span**$(\mathbf{u}_1, \ldots, \mathbf{u}_r)^{\perp}$.

Explicit choices of Σ/L we consider

- We generally stick to choices of Σ like $\Sigma = \sum_{k=1}^{p} \varphi_k \mathbf{v}_k \mathbf{v}_k^T$ which result in a transform L which changes the singular values of X.

Explicit choices of Σ/L we consider

- We generally stick to choices of Σ like $\Sigma = \sum_{k=1}^{p} \varphi_k \mathbf{v}_k \mathbf{v}_k^T$ which result in a transform L which changes the singular values of X.
- The following choices of Σ yield and L which corresponds exactly to the trim and lava transforms we have discussed:
    - The trim transform is obtained for the choice

$$\varphi_k = \begin{cases} \frac{1}{\tau^2} - \frac{1}{\lambda_k} & \text{if } \lambda_k > \tau^2, \\ 0 & \text{if } \lambda_k \leq \tau^2 . \end{cases}$$

Explicit choices of Σ/L we consider

- We generally stick to choices of Σ like $\Sigma = \sum_{k=1}^{p} \varphi_k \mathbf{v}_k \mathbf{v}_k^T$ which result in a transform L which changes the singular values of X.
- The following choices of Σ yield and L which corresponds exactly to the trim and lava transforms we have discussed:
  - The trim transform is obtained for the choice

$$\varphi_k = \begin{cases} \frac{1}{\tau^2} - \frac{1}{\lambda_k} & \text{if } \lambda_k > \tau^2, \\ 0 & \text{if } \lambda_k \leq \tau^2 . \end{cases}$$

  - The lava transform is obtained for the choice

$$\varphi_k = \frac{1}{n\mu_2} \text{ for all } k,$$

so that $\Sigma = I_p/(n\mu_2)$.

Example: Gaussian Process Regression

- Recalling the form of the marginal posterior in general

$$\Pi_f(F \mid X, Y) = \frac{\int_F e^{-\frac{1}{2\sigma_\varepsilon^2}\|L(Y - f(X))\|_2^2} d\Pi_f(f)}{\int e^{-\frac{1}{2\sigma_\varepsilon^2}\|L(Y - f(X))\|_2^2} d\Pi_f(f)}.$$

- Place a GP prior on: $f \sim GP(0, k)$ for some positive definite kernel k.

Example: Gaussian Process Regression

- Recalling the form of the marginal posterior in general

$$\Pi_f(F \mid X, Y) = \frac{\int_F e^{-\frac{1}{2\sigma_\varepsilon^2}\|L(Y-f(X))\|_2^2} d\Pi_f(f)}{\int e^{-\frac{1}{2\sigma_\varepsilon^2}\|L(Y-f(X))\|_2^2} d\Pi_f(f)}.$$

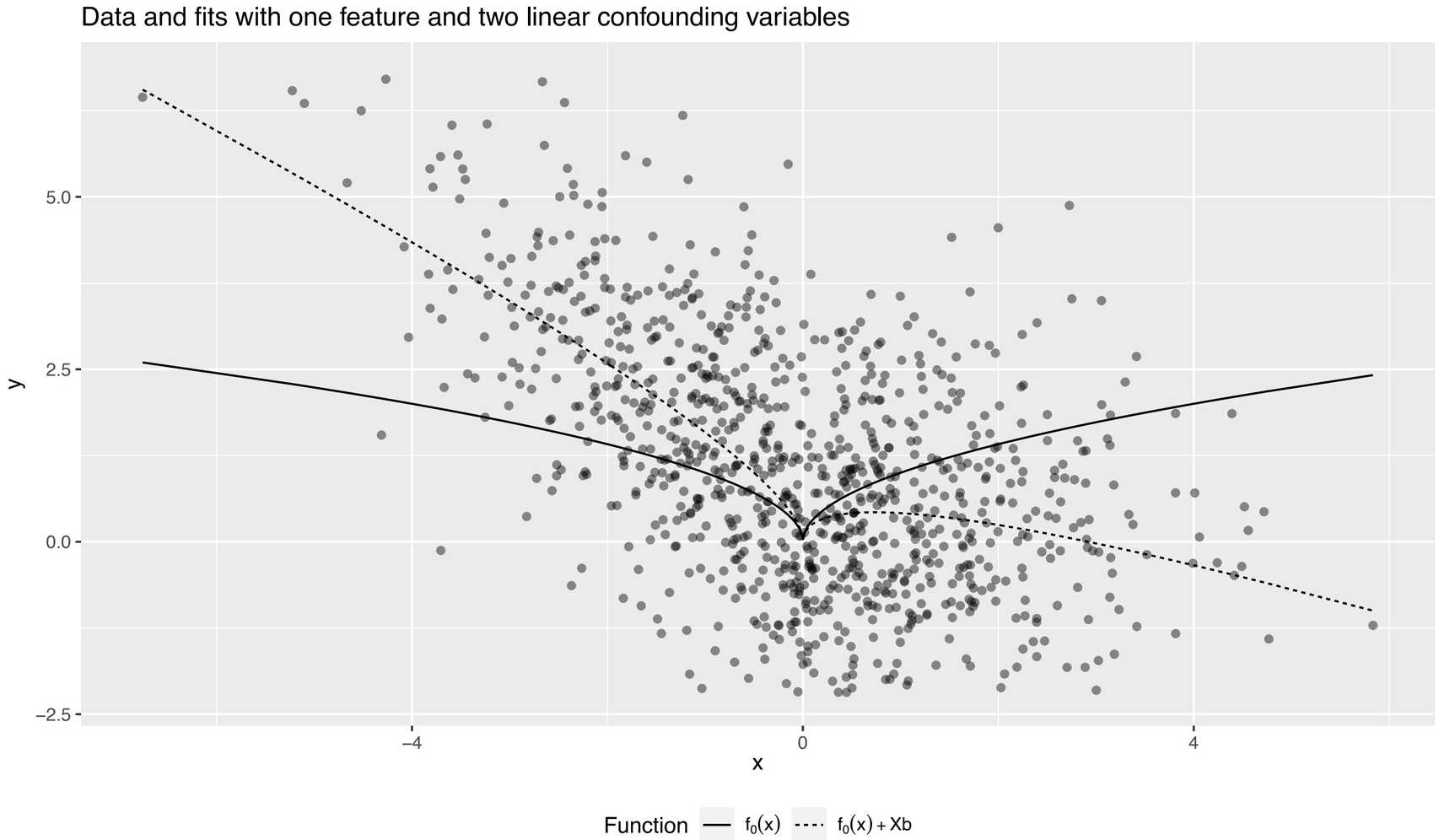
- Place a GP prior on: $f \sim GP(0, k)$ for some positive definite kernel k.

- The marginal posterior is a GP with mean and covariance function given

$$\mu_p(\mathbf{x}) = \mathbf{k}_n(\mathbf{x})^T (K_{X,X} + \sigma_\varepsilon^2 I_n + X\Sigma X^T)^{-1}\mathbf{y}$$

$$k_p(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}, \mathbf{x}') - \mathbf{k}_n(\mathbf{x})^T (K_{X,X} + \sigma_\varepsilon^2 I_n + X\Sigma X^T)^{-1}\mathbf{k}_n(\mathbf{x}')$$

Example: Gaussian Process Regression



Left: Dataset generated according to

$$X = H\Gamma + E, \qquad H_{ij}, \Gamma_{ij}, E_{ij} \sim \mathcal{N}(0,1)$$

$$\delta = (0.57, -1.35)^T$$

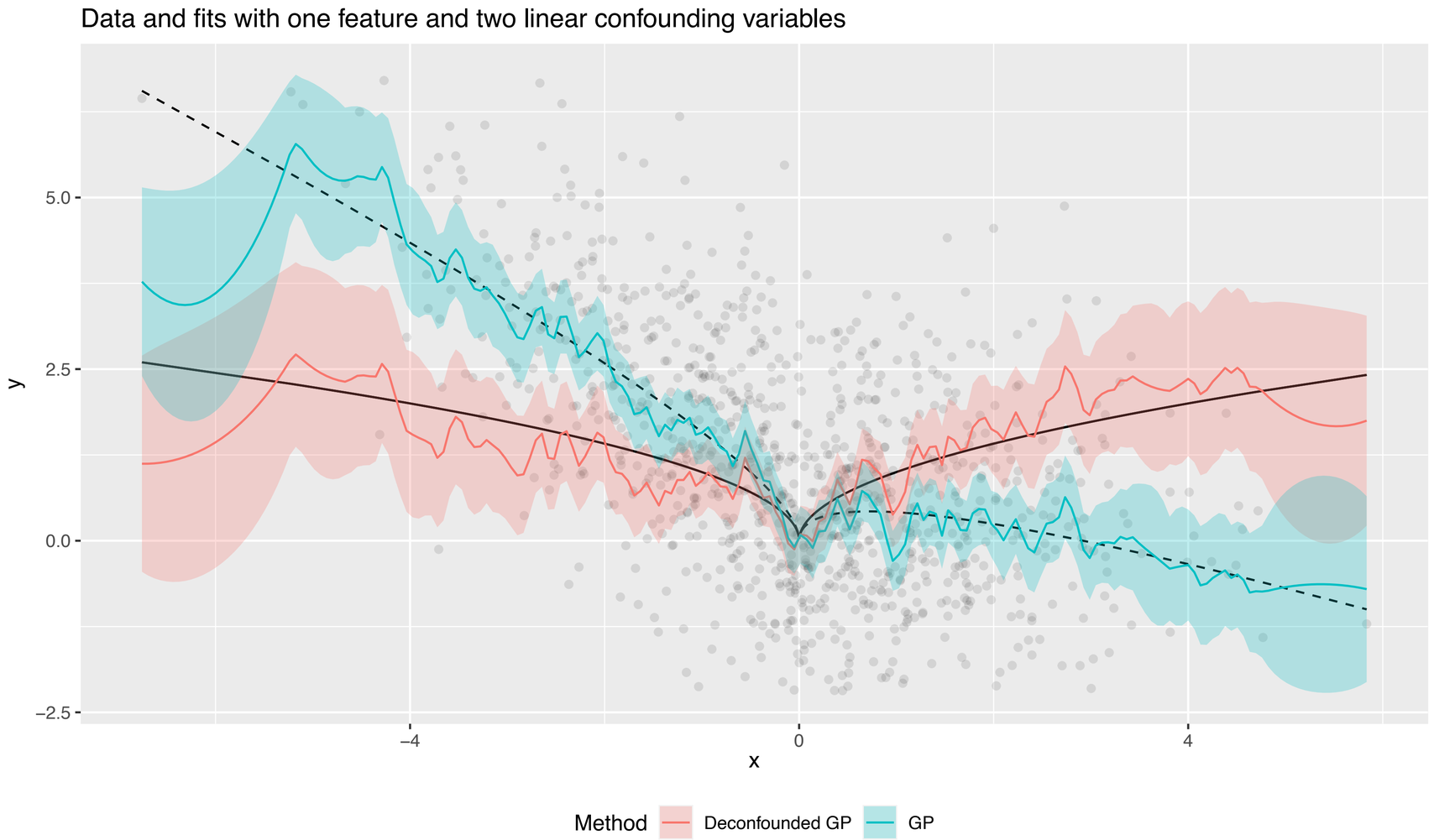
$$Y = \sqrt{|X|} + H\delta + \epsilon$$

$$(n, p, q) = (1000, 1, 2)$$

$$\Pi(f) = GP(0, k_{Mat-1/2}), \Pi(b) = \mathcal{N}(0, 1/n)$$

IMPERIAL

Example: Gaussian Process Regression



Left: Dataset generated according to

$$X = H\Gamma + E, \qquad H_{ij}, \Gamma_{ij}, E_{ij} \sim \mathcal{N}(0,1)$$

$$\delta = (0.57, -1.35)^T$$

$$Y = \sqrt{|X|} + H\delta + \epsilon$$

$$(n, p, q) = (1000, 1, 2)$$

$$\Pi(f) = GP(0, k_{Mat-1/2}), \Pi(b) = \mathcal{N}(0, 1/n)$$

IMPERIAL

Empirical results: Gaussian Process Regression

		(Number of features, Number of confounders)			
Metric	Method	$(1, 2)$	$(4, 2)$	$(8, 2)$	$(8, 4)$
L_1	GP	0.695 ± 0.481	0.490 ± 0.228	0.355 ± 0.140	0.454 ± 0.140
	Decon-GP	$\mathbf{0.285 \pm 0.190}$	$\mathbf{0.268 \pm 0.072}$	$\mathbf{0.237 \pm 0.054}$	$\mathbf{0.292 \pm 0.093}$
L_2	GP	0.752 ± 0.951	0.419 ± 0.402	0.234 ± 0.191	0.353 ± 0.220
	Decon-GP	$\mathbf{0.139 \pm 0.172}$	$\mathbf{0.118 \pm 0.064}$	$\mathbf{0.088 \pm 0.037}$	$\mathbf{0.127 \pm 0.071}$
Coverage	GP	0.445 ± 0.437	0.964 ± 0.081	0.996 ± 0.011	0.994 ± 0.014
	Decon-GP	$\mathbf{0.940 \pm 0.192}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
Length	GP	1.057 ± 0.098	2.963 ± 0.178	3.619 ± 0.061	3.747 ± 0.055
	Decon-GP	1.460 ± 0.301	3.012 ± 0.177	3.637 ± 0.059	3.753 ± 0.053
Time	GP	2.888 ± 0.113	3.241 ± 0.129	4.074 ± 0.218	4.499 ± 0.312
	Decon-GP	2.958 ± 0.181	3.202 ± 0.151	4.173 ± 0.277	4.480 ± 0.246

Table 1: Estimates of the different metrics for the various GP regression methods where there is linear confounding in the data. In each case $n = 1000$, $E_{ij}, H_{ij}, \Gamma_{ij}, \delta_i \sim \mathcal{N}(0, 1)$ and $X = H\Gamma + E$. Metrics are computed at points where $\|\mathbf{x}_*\|_2 = 1$, where the size of the perturbation is constant.

IMPERIAL

Example: Sparse Linear Regression

- Consider the case that $f(X) = X\beta$ for some sparse β. Place a prior distribution on f by placing a prior on the coefficients β.

Chernozhukov, V., Hansen, C., Liao, Y.. *A lava attack on the recovery of sums of dense and sparse signals.* AOS, 2017.

Castillo, I., Schmidt-Hieber, J., van der Vaart, A.. *Bayesian linear regression with sparse priors.* AOS, 2015.

Example: Sparse Linear Regression

- Consider the case that $f(X) = X\beta$ for some sparse β. Place a prior distribution on f by placing a prior on the coefficients β.
- Famous choice (in unconfounded setting) is the ‘Bayesian Lasso’, the resulting mode is the same as the estimator proposed in Chernozukhov et al. (2017).

Chernozhukov, V., Hansen, C., Liao, Y.. *A lava attack on the recovery of sums of dense and sparse signals.* AOS, 2017.

Castillo, I., Schmidt-Hieber, J., van der Vaart, A.. *Bayesian linear regression with sparse priors.* AOS, 2015.

Example: Sparse Linear Regression

- Consider the case that $f(X) = X\beta$ for some sparse β. Place a prior distribution on f by placing a prior on the coefficients β.
- Famous choice (in unconfounded setting) is the ‘Bayesian Lasso’, the resulting mode is the same as the estimator proposed in Chernozukhov et al. (2017).
- However Castillo et al. (2015) show that this is a suboptimal choice for a prior on β in the case of no confounding.

Chernozhukov, V., Hansen, C., Liao, Y.. *A lava attack on the recovery of sums of dense and sparse signals.* AOS, 2017.

Castillo, I., Schmidt-Hieber, J., van der Vaart, A.. *Bayesian linear regression with sparse priors.* AOS, 2015.

Example: Sparse Linear Regression

- Consider the case that $f(X) = X\beta$ for some sparse β. Place a prior distribution on f by placing a prior on the coefficients β.
- Famous choice (in unconfounded setting) is the ‘Bayesian Lasso’, the resulting mode is the same as the estimator proposed in Chernozukhov et al. (2017).
- However Castillo et al. (2015) show that this is a suboptimal choice for a prior on β in the case of no confounding.
- The model selection prior considered in Castillo et al. (2015) is a better choice.

Chernozhukov, V., Hansen, C., Liao, Y.. *A lava attack on the recovery of sums of dense and sparse signals.* AOS, 2017.

Castillo, I., Schmidt-Hieber, J., van der Vaart, A.. *Bayesian linear regression with sparse priors.* AOS, 2015.

Example: Sparse Linear Regression

- Consider the case that $f(X) = X\beta$ for some sparse β. Place a prior distribution on f by placing a prior on the coefficients β.
- Famous choice (in unconfounded setting) is the 'Bayesian Lasso', the resulting mode is the same as the estimator proposed in Chernozukhov et al. (2017).
- However Castillo et al. (2015) show that this is a suboptimal choice for a prior on β in the case of no confounding.
- The model selection prior considered in Castillo et al. (2015) is a better choice.
- It is given as follows:

1. The sparsity s of β is drawn according to some π_p a distribution on $\{1,\ldots,p\}$. E.g. $\pi_p = \text{Binomial}(p, \alpha)$, for some probability of inclusion α.
2. The active set S given $|S| = s$ of β is drawn uniformly from the $\binom{p}{s}$ subsets of $\{1,\ldots,p\}$ of size s.
3. $\beta_i | S \overset{ind}{\sim} \begin{cases} \text{Lap}(\lambda), & i \in S, \\ \delta_0, & i \notin S. \end{cases}$

Chernozhukov, V., Hansen, C., Liao, Y.. *A lava attack on the recovery of sums of dense and sparse signals.* AOS, 2017.

Castillo, I., Schmidt-Hieber, J., van der Vaart, A.. *Bayesian linear regression with sparse priors.* AOS, 2015.

Example: Sparse Linear Regression

- Our marginal posterior distribution on β is given by

$$\Pi_\beta(B \mid X, Y) = \frac{\int_B e^{-\frac{1}{2\sigma_\varepsilon^2}\|LY - LX\beta\|_2^2} d\Pi_{MS}(\beta)}{\int e^{-\frac{1}{2\sigma_\varepsilon^2}\|LY - LX\beta\|_2^2} d\Pi_{MS}(\beta)} = \frac{\int_B e^{-\frac{1}{2\sigma_\varepsilon^2}\|\tilde{Y} - \tilde{X}\beta\|_2^2} d\Pi_{MS}(\beta)}{\int e^{-\frac{1}{2\sigma_\varepsilon^2}\|\tilde{Y} - \tilde{X}\beta\|_2^2} d\Pi_{MS}(\beta)},$$

  for $\tilde{Y} := LY$ and $\tilde{X} := LX$.

Example: Sparse Linear Regression

- Our marginal posterior distribution on β is given by

$$\Pi_\beta(B \mid X, Y) = \frac{\int_B e^{-\frac{1}{2\sigma_\varepsilon^2}\|LY - LX\beta\|_2^2} d\Pi_{MS}(\beta)}{\int e^{-\frac{1}{2\sigma_\varepsilon^2}\|LY - LX\beta\|_2^2} d\Pi_{MS}(\beta)} = \frac{\int_B e^{-\frac{1}{2\sigma_\varepsilon^2}\|\tilde{Y} - \tilde{X}\beta\|_2^2} d\Pi_{MS}(\beta)}{\int e^{-\frac{1}{2\sigma_\varepsilon^2}\|\tilde{Y} - \tilde{X}\beta\|_2^2} d\Pi_{MS}(\beta)},$$

  for $\tilde{Y} := LY$ and $\tilde{X} := LX$.

- No analytic form for the marginal posterior.

Ray, K., Szabó, B.. *Variational Bayes for High-Dimensional Linear Regression with Sparse Priors.* JASA, 2022.

Example: Sparse Linear Regression

- Our marginal posterior distribution on β is given by

$$\Pi_\beta(B \mid X, Y) = \frac{\int_B e^{-\frac{1}{2\sigma_\varepsilon^2}\|LY - LX\beta\|_2^2} d\Pi_{MS}(\beta)}{\int e^{-\frac{1}{2\sigma_\varepsilon^2}\|LY - LX\beta\|_2^2} d\Pi_{MS}(\beta)} = \frac{\int_B e^{-\frac{1}{2\sigma_\varepsilon^2}\|\tilde{Y} - \tilde{X}\beta\|_2^2} d\Pi_{MS}(\beta)}{\int e^{-\frac{1}{2\sigma_\varepsilon^2}\|\tilde{Y} - \tilde{X}\beta\|_2^2} d\Pi_{MS}(\beta)},$$

  for $\tilde{Y} := LY$ and $\tilde{X} := LX$.

- No analytic form for the marginal posterior.

- Instead one must sample from the posterior using MCMC or approximate it using variational inference (e.g. as per Ray and Szabo, 2022).

Ray, K., Szabó, B.. *Variational Bayes for High-Dimensional Linear Regression with Sparse Priors.* JASA, 2022.

Example: Sparse Linear Regression

- Our marginal posterior distribution on β is given by

$$\Pi_\beta(B \mid X, Y) = \frac{\int_B e^{-\frac{1}{2\sigma_\varepsilon^2}\|LY - LX\beta\|_2^2} d\Pi_{MS}(\beta)}{\int e^{-\frac{1}{2\sigma_\varepsilon^2}\|LY - LX\beta\|_2^2} d\Pi_{MS}(\beta)} = \frac{\int_B e^{-\frac{1}{2\sigma_\varepsilon^2}\|\tilde{Y} - \tilde{X}\beta\|_2^2} d\Pi_{MS}(\beta)}{\int e^{-\frac{1}{2\sigma_\varepsilon^2}\|\tilde{Y} - \tilde{X}\beta\|_2^2} d\Pi_{MS}(\beta)},$$

  for $\tilde{Y} := LY$ and $\tilde{X} := LX$.

- No analytic form for the marginal posterior.

- Instead one must sample from the posterior using MCMC or approximate it using variational inference (e.g. as per Ray and Szabo, 2022).

- However the 'nice' thing is that one can just supply the transformed data $(\tilde{X}, \tilde{Y})$ to whichever procedure one would use in place of (X, Y) — the deconfounding amounts to a preprocessing step.

Ray, K., Szabó, B.. *Variational Bayes for High-Dimensional Linear Regression with Sparse Priors.* JASA, 2022.

Empirical results: High-dimensional linear regression

Method	FreqDec		BayesDec		SAS
Transform	Trim	Lava	Trim	Lava	N/A
ℓ_2-error	0.637 ± 0.138	0.648 ± 0.120	$\mathbf{0.256 \pm 0.091}$	0.296 ± 0.075	0.817 ± 0.501
ℓ_1-error	2.082 ± 0.842	1.949 ± 0.712	$\mathbf{0.483 \pm 0.160}$	0.565 ± 0.230	3.139 ± 1.223
Precision	0.291 ± 0.082	0.338 ± 0.088	0.995 ± 0.014	$\mathbf{1.000 \pm 0.000}$	0.758 ± 0.170
Recall	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
F_1	0.216 ± 0.082	0.243 ± 0.085	0.499 ± 0.004	$\mathbf{0.500 \pm 0.000}$	0.431 ± 0.069

Table 4: Estimates of the relevant metrics over 100 realisations of the data described in the text. (n, p, s_0, q) are given by $(100, 200, 5, 3)$.

Theoretical Results

Confounded Model Assumptions

- In the confounded model, we prove our results under P_0 given by

$$Y = f_0(X) + H\delta^0 + \nu,$$

with some $\delta^0 \in \mathbb{R}^q$, a function f_0 for which we may specify a form, and mean-zero, sub-Gaussian noise ν.

Confounded Model Assumptions

- In the confounded model, we prove our results under P_0 given by

$$Y = f_0(X) + H\delta^0 + \nu,$$

  with some $\delta^0 \in \mathbb{R}^q$, a function f_0 for which we may specify a form, and mean-zero, sub-Gaussian noise ν.

- The rows of X and H are assumed sub-Gaussian with $\Gamma = \text{cov}(H, X)$, and $X = H\Gamma + E$ for some E with $\text{cov}(E) = \Sigma_E$.

Confounded Model Assumptions

- In the confounded model, we prove our results under P_0 given by

$$Y = f_0(X) + H\delta^0 + \nu,$$

  with some $\delta^0 \in \mathbb{R}^q$, a function f_0 for which we may specify a form, and mean-zero, sub-Gaussian noise ν.

- The rows of X and H are assumed sub-Gaussian with $\Gamma = \text{cov}(H, X)$, and $X = H\Gamma + E$ for some E with $\text{cov}(E) = \Sigma_E$.

- Suppose that $\|\delta^0\|_2 = \mathcal{O}(1)$ and $\sigma_{\min}(\Gamma) = \Omega(1)$, which ensures

  1. ε in the perturbed version of the confounded model has finite variance proxy.

  2. The confounding is ‘dense’, which ensures that the latent b lies mostly in the first singular directions of X and that the prior can shrink the perturbation effectively.

Nonparametric regression

- For $X \sim q$ a density on a compact space (e.g. $[0,1]^p$), we obtain a contraction rate in terms of the norm

$$\|f\|_{q,2}^2 = \int_{[0,1]^p} f^2(\mathbf{x}) q(\mathbf{x}) d\mathbf{x}$$

- We assume here $\sigma_{\max}(X\Sigma^{1/2}) = \mathcal{O}_p(1)$, though this can be weakened.

Nonparametric regression

Theorem: Gaussian process prior in confounded setting

Let P_0 be determined by $Y = f_0(X) + H\delta^0 + \nu$ satisfying the confounded model assumptions with f_0 which is $\beta-$Hölder.

Let $X \sim q$ with support $[0,1]^p$ and suppose that the GP prior on f places probability 1 on $C^\alpha([0,1]^p)$ and satisfies $\varphi_{f_0}(\varepsilon_n) \leq n\varepsilon_n^2$.

Suppose that the prior on b satisfies (via L), $\sigma_{\max}(LX) = \mathcal{O}_p(1)$ and $\min(\alpha, \beta) > p/2$.

Nonparametric regression

Theorem: Gaussian process prior in confounded setting

Let P_0 be determined by $Y = f_0(X) + H\delta^0 + \nu$ satisfying the confounded model assumptions with f_0 which is $\beta-$Hölder.

Let $X \sim q$ with support $[0,1]^p$ and suppose that the GP prior on f places probability 1 on $C^\alpha([0,1]^p)$ and satisfies $\varphi_{f_0}(\varepsilon_n) \leq n\varepsilon_n^2$.

Suppose that the prior on b satisfies (via L), $\sigma_{\max}(LX) = \mathcal{O}_p(1)$ and $\min(\alpha, \beta) > p/2$.

Then

$$E_0\left[\Pi_f\left(f : \|f - f_0\|_{q,2} > M\varepsilon_n \,|\, X, Y\right)\right] \to 0$$

- Generally gives rate $\varepsilon_n \asymp n^{-\min(\alpha,\beta)/(2\alpha+p)}$ for a $\beta-$smooth f_0 e.g. in the case of sub-Gaussian X with the trim or lava transforms if $\min(\alpha, \beta) > p/2$.
- Also have a (slower) rate when $\min(\alpha, \beta) \leq p/2$, but presentation of this requires quite a lot of further discussion.

Summary

- Introduced a general Bayesian method for approaching linearly confounded models.
- Described how to use it in the specific case of Gaussian process regression and high-dimensional linear regression.

Summary

- Introduced a general Bayesian method for approaching linearly confounded models.
- Described how to use it in the specific case of Gaussian process regression and high-dimensional linear regression.

In our work we also

- Present further and more refined theoretical results.
- Introduce a similar approach in the case of a non-linear perturbation, where the idea is to place a GP prior on $g(X)$ instead of a linear prior on Xb.

References

Ćevid, D., Bühlmann, P., Meinhausen, N. *Spectral deconfounding via perturbed sparse linear models.* In: JMLR, 2020.

Chernozhukov, V., Hansen, C., Liao, Y.. *A lava attack on the recovery of sums of dense and sparse signals.* In: AOS, 2017

Castillo, I., Schmidt-Hieber, J., van der Vaart, A.. *Bayesian linear regression with sparse priors.* In: AOS, 2015.

Ray, K., Szabó, B.. *Variational Bayes for High-Dimensional Linear Regression with Sparse Priors.* In: JASA, 2022.

van der Vaart, A., van Zanten, H.. *Information Rates of Nonparametric Gaussian Process Methods.* In: JMLR, 2011

- Code available at: https://github.com/lukemmtravis/BayesianDeconfounding/
- Paper to follow on arXiv shortly.

IMPERIAL

Appendix: Low-dimensional perturbed linear regression

- Consider with $p < n$ the model:

$$Y = X\beta + Xb + \varepsilon.$$

- Placing $\Pi(\beta) = \mathcal{N}_p(0, \Sigma_\beta)$ and $\Pi(b) = \mathcal{N}_p(0, \Sigma_b)$, marginal posterior of β given by

$$\tilde{\Pi}_\beta(\beta \,|\, X, Y) = \mathcal{N}_p\left(\underbrace{[\tilde{X}^T\tilde{X} + \Sigma_\beta^{-1}]^{-1}\tilde{X}^T\tilde{Y}}_{\tilde{\mu}_\beta},\ [\tilde{X}^T\tilde{X} + \Sigma_\beta^{-1}]^{-1} \right),$$

  where $\tilde{X} = LX$, $\tilde{Y} = LY$ for L given by

$$L^TL = I - X(X^TX + \Sigma_b^{-1})^{-1}X^T$$

- It is the same as the normal posterior with no perturbation in the likelihood, but with $(\tilde{X}, \tilde{Y})$ in place of (X, Y).

Appendix: Low-dimensional perturbed linear regression

- For any choice of Σ_β, and $W = X\Sigma_\beta^{1/2} = \sum_{k=1}^{p} \sqrt{\lambda_k}\mathbf{u}_k\mathbf{v}_k^T$ the ordinary posterior has bias given by

$$E_0(\mu_\beta - \beta^0) = \Sigma_\beta^{1/2}\left[-\sum_{k=1}^{p}\frac{1}{1+\lambda_k}\mathbf{v}_k\mathbf{v}_k^T\right]\Sigma_\beta^{-1/2}\beta^0 + \Sigma_\beta^{1/2}\left[\sum_{k=1}^{p}\frac{\lambda_k}{1+\lambda_k}\mathbf{v}_k\mathbf{v}_k^T\right]\Sigma_\beta^{-1/2}b^0$$

  where typically one wants $\lambda_k \to \infty$ to kill the first term in the case of no confounding. This however leaves a bias of b^0.

- If one chooses $\Sigma_b = \left[\sum_{k=1}^{p}\varphi_k^{-1}[\Sigma_\beta^{1/2}\mathbf{v}_k][\mathbf{v}_k^T\Sigma_\beta^{1/2}]\right]$, then the bias from the marginal posterior with the deconfounding prior is given

$$E_0(\tilde{\mu}_\beta - \beta^0) = \Sigma_\beta^{1/2}\left[-\sum_{k=1}^{p}\frac{1}{1+\nu_k}\mathbf{v}_k\mathbf{v}_k^T\right]\Sigma_\beta^{-1/2}\beta^0 + \Sigma_\beta^{1/2}\left[\sum_{k=1}^{p}\frac{\nu_k}{1+\nu_k}\mathbf{v}_k\mathbf{v}_k^T\right]\Sigma_\beta^{-1/2}b^0,$$

where now $\nu_k = \lambda_k\varphi_k/(\lambda_k+\varphi_k)$, such that $0 \leq \nu_k \leq \lambda_k -$ and one can control which of the terms to kill via the choice of φ_k.

Theoretical Results

Confounded Model Assumptions

- In the confounded model, we prove our results under P_0 given by

$$Y = f_0(X) + H\delta^0 + \nu,$$

  with some $\delta^0 \in \mathbb{R}^q$, a function f_0 for which we may specify a form, and mean-zero, sub-Gaussian noise ν.

- The rows of X and H are assumed sub-Gaussian with $\Gamma = \text{cov}(H, X)$, and $X = H\Gamma + E$ for some E with $\text{cov}(E) = \Sigma_E$.

- Suppose that $\|\delta^0\|_2 = \mathcal{O}(\sigma_{\min}(\Gamma))$ and $\sigma_{\min}(\Gamma) = \Omega(\sqrt{p})$, which ensures

  1. ε in the perturbed version of the confounded model has finite variance proxy.

  2. The confounding is ‘dense’, which ensures that the latent b lies mostly in the first singular directions of X and that the prior can shrink the perturbation effectively.

Goal and structure of results

- Ultimate goal is to prove contraction rates for the marginal posterior. I.e. for some $\varepsilon_n \to 0$

$$E_0 \left[\Pi_f \left(f : \|f - f_0\| > M\varepsilon_n \,|\, X, Y \right) \right] \to 0,$$

  for some appropriate $\|\cdot\|$ norm, where E_0 is the expectation under P_0.

- First prove the above for the **perturbed model**

$$Y = f_0(X) + Xb^0 + \varepsilon$$

  before showing that the necessary assumptions are verified under the **confounded model**

$$Y = f_0(X) + H\delta^0 + \nu$$

  under the assumptions on the previous slide.

Nonparametric regression

- We begin with nonparametric regression.
- For a fixed design $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ nonparametric regression, one typically consider contraction rates in the norm

$$\|f\|_{n,2}^2 := \frac{1}{n}\sum_{i=1}^{n} f^2(\mathbf{x}_i) = \frac{1}{n}\mathbf{f}^T\mathbf{f},$$

  we thus usually get contraction rates in this empirical norm.
- Instead we work with

$$\|f\|_{L,n,2}^2 = \frac{1}{n}\mathbf{f}^T(L^TL)\mathbf{f},$$

  which one can think of as an empirical norm induced by heterogenous errors $\varepsilon \sim \mathcal{N}_n(0,(L^TL)^{-1})$.
- For now just know that this norm is strictly weaker — $\|\cdot\|_{L,n,2} \leq \|\cdot\|_{n,2}$.

Nonparametric regression

Theorem: Fixed design regression

Let P_0 be determined by $Y = f_0(X) + Xb^0 + \varepsilon$, with ε sub-Gaussian (**not** necessarily independent of X) with finite variance proxy σ_ε^2 for any X.

Suppose that the prior on b satisfies $\|LXb^0\|_2 = o(1)$.

Suppose there exists sets $F_n \subset \mathcal{F}$ and constants $D_1, D_2 > 0$ such that

$$\log N(\varepsilon_n, F_n, \|\cdot\|_{L,n,2}) \leq D_1 n\varepsilon_n^2$$

$$\Pi(F_n^C) \leq e^{-(D_2+4)n\varepsilon_n^2}$$

$$\Pi_f(f : \|f - f_0\|_{n,2} < \varepsilon_n/2) \geq e^{-D_2 n\varepsilon_n^2}.$$

Then

$$E_0\left[\Pi_f\left(f : \|f - f_0\|_{L,n,2} > M\varepsilon_n \,|\, Y\right) | X\right] \to 0.$$

The prior on b needs to 'shrink' LXb^0 sufficiently.

Weaker than normal condition
Weaker than normal norm for the rate

Even if we assume the covering number with the stronger norm, we do not necessarily get contraction in the stronger norm as we can only test with the $\|\cdot\|_{L,n,2}$ norm.

IMPERIAL

Comparison of norms

- For $W := X\Sigma^{1/2} = \sum_{k=1}^{p} \sqrt{\lambda_k^W}(\mathbf{u}_k^W)(\mathbf{v}_k^W)^T$,

$$L^T L = \sum_{k=1}^{p} \frac{1}{1+\lambda_k^W}(\mathbf{u}_k^W)(\mathbf{u}_k^W)^T + \sum_{k=p+1}^{n} (\mathbf{u}_k^W)(\mathbf{u}_k^W)^T$$

- If we have $\sigma_{\max}(W) = \mathcal{O}(\sqrt{M_n})$ for some deterministic sequence M_n, then have the relationship

$$\frac{1}{1+M_n}\|f\|_{n,2}^2 \leq \|f\|_{L,n,2}^2 \leq \|f\|_{n,2}^2,$$

  with the lower bound achieved by some function if there exists k with $\lambda_k^W = M_n$.

- For the main priors we consider (lava and trim), M_n can be bounded by a constant and the norms in this case are equivalent.
- E.g. for the lava transform and sub-Gaussian X, $M_n = 1 + \log n/n + Cp/n < C'$ on an event of probability tending to 1 for some constants C and C'.
- However in our later results we allow for the possibility that M_n possibly goes to infinity.

Nonparametric regression

Theorem: Gaussian process prior with fixed design

Let P_0 be determined by $Y = f_0(X) + Xb^0 + \varepsilon$, with ε sub-Gaussian with finite variance proxy σ_ε^2 for any X.

Suppose that the prior on b satisfies $\|LXb^0\|_2 = o(1)$.

Let the prior on f be a GP whose concentration function satisfies $\varphi_{f_0}(\varepsilon_n) \leq n\varepsilon_n^2$.

Then

$$E_0\left[\Pi_f\left(f : \|f - f_0\|_{L,n,2} > M\varepsilon_n \,|\, Y\right) | X\right] \to 0.$$

- Note, in the case of the trim and lava transforms, $\|\cdot\|_{L,n,2}$ can be replaced by $\|\cdot\|_{n,2}$

IMPERIAL

Nonparametric regression

- To extend to the confounded setting, it makes more sense to think about a random design.
- We will assume that X is distributed according to some known density q on a compact set, $[0,1]^p$ say, and prove a result in terms of the norm

$$\|f\|_{q,2}^2 = \int_{[0,1]^p} f^2(\mathbf{x}) q(\mathbf{x}) d\mathbf{x}.$$

- We can prove a useful high-probability relationship between the $\|\cdot\|_{n,2}$ norm and the $\|\cdot\|_{q,2}$ norm, but we use the preceding results which are in terms of the $\|\cdot\|_{L,n,2}$ norm.
- As such, we need to make the assumption which allows us to compare the $\|\cdot\|_{L,n,2}$ norm and the $\|\cdot\|_{n,2}$ norm — i.e. $\sigma_{\max}(X\Sigma^{1/2}) = \mathcal{O}_p(\sqrt{M_n})$.

Nonparametric regression

Theorem: Gaussian process prior with random design

Let P_0 be determined by $Y = f_0(X) + Xb^0 + \varepsilon$, with ε sub-Gaussian with finite variance proxy σ_ε^2 for any X.

Suppose that the prior on b satisfies $\|LXb^0\|_2 = o_p(1)$.

Assume additionally that $X \sim q$ with support $[0,1]^p$, and for some $\alpha > 0$ the GP prior on f gives probability one to the Hölder space $C^\alpha\left([0,1]^p\right)$ with $\varphi_{f_0}(\varepsilon_n) \leq n\varepsilon_n^2$.

Assume that $\sigma_{\max}(X\Sigma^{1/2}) = \mathcal{O}_p(\sqrt{M_n})$ for some sequence M_n and define $t_n^2 = 1/(1+M_n)$.

(Reminder: then $t_n\|f\|_{n,2} \leq \|f\|_{L,n,2} \leq \|f\|_{n,2}$ on such an event of probability 1)

Then for sufficiently large M

$$E_0\left[\Pi_f\left(f : \|f - f_0\|_{q,2} > M\varepsilon_n/t_n \,|\, X, Y\right)\right] \to 0, \qquad \text{if } \varepsilon_n \leq n^{-\frac{p/2}{2\alpha+p}} t_n^{-\frac{p}{2\alpha+p}}$$

$$E_0\left[\Pi_f\left(f : \|f - f_0\|_{q,2} > Mn^{1/2}\varepsilon_n^{(2\alpha+2p)/p} \,|\, X, Y\right)\right] \to 0, \qquad \text{if } \varepsilon_n \geq n^{-\frac{p/2}{2\alpha+p}} t_n^{-\frac{p}{2\alpha+p}}$$

Nonparametric regression

$$E_0\left[\Pi_f\left(f:\|f-f_0\|_{q,2}>M\varepsilon_n/t_n\,|\,X,Y\right)\right]\to 0, \qquad \text{if } \varepsilon_n \leq n^{-\frac{p/2}{2\alpha+p}} t_n^{-\frac{p}{2\alpha+p}}$$

$$E_0\left[\Pi_f\left(f:\|f-f_0\|_{q,2}>Mn^{1/2}\varepsilon_n^{(2\alpha+2p)/p}\,|\,X,Y\right)\right]\to 0, \qquad \text{if } \varepsilon_n \geq n^{-\frac{p/2}{2\alpha+p}} t_n^{-\frac{p}{2\alpha+p}}$$

- For commonly used GP priors and a $\beta-$smooth f_0, the condition $\varphi(\varepsilon_n) \leq n\varepsilon_n^2$ is verified for $\varepsilon_n \asymp n^{-\frac{\min(\alpha,\beta)}{2\alpha+p}}$.
- For $t_n \asymp 1$, first case thus occurs with a if $\min(\alpha,\beta) > p/2$ and gives the usual rate.
- For $t_n \asymp 1$, the second case occurs if $\min(\alpha,\beta) \leq p/2$ and gives a slower 'rate' which results from the $\|\cdot\|_{q,2}$ norm extrapolating from the design to the support of q. Known problem in non-confounded setting (e.g. van der Vaart and van Zanten, 2011).

van der Vaart, A., van Zanten, H.. *Information rates of nonparametric Gaussian process methods.* JMLR, 2011

Nonparametric regression

Theorem: Gaussian process prior in confounded setting

Let P_0 be determined by $Y = f_0(X) + H\delta^0 + \nu$ satisfying the confounded model assumptions.

Let $X \sim q$ with support $[0,1]^p$ and that the GP prior on f places probability 1 on $C^\alpha([0,1]^p)$ and satisfies $\varphi_{f_0}(\varepsilon_n) \leq n\varepsilon_n^2$.

Suppose that the prior on b satisfies (via L), $\sigma_{\max}(LX) = \mathcal{O}_p(\sqrt{p})$ and $\sigma_{\max}(X\Sigma^{1/2}) = \mathcal{O}_p(\sqrt{M_n})$ for some sequence M_n possibly going to ∞ and define $t_n^2 = 1/(1+M_n)$.

Then

$$E_0\left[\Pi_f\left(f : \|f - f_0\|_{q,2} > M\varepsilon_n/t_n \,|\, X, Y\right)\right] \to 0, \qquad \text{if } \varepsilon_n \leq n^{-\frac{p/2}{2\alpha+p}} t_n^{-\frac{p}{2\alpha+p}}$$

$$E_0\left[\Pi_f\left(f : \|f - f_0\|_{q,2} > Mn^{1/2}\varepsilon_n^{(2\alpha+2p)/p} \,|\, X, Y\right)\right] \to 0, \qquad \text{if } \varepsilon_n \geq n^{-\frac{p/2}{2\alpha+p}} t_n^{-\frac{p}{2\alpha+p}}$$

- Gives rate $\varepsilon_n \asymp n^{-\min(\alpha,\beta)/(2\alpha+p)}$ for a $\beta-$smooth f_0 e.g. in the case of sub-Gaussian X with the trim or lava transforms if $\min(\alpha, \beta) > p/2$.

IMPERIAL

High-dimensional linear regression

- Now consider the high-dimensional ($n < p$) setting.
- A linear function $f_0(X) = X\beta^0$, where for some $0 \leq s_0 \ll p$,

$$s_0 = |\{i : \beta_i^0 \neq 0\}|.$$

- We first consider the perturbed model with a fixed design, i.e. a P_0 under which

$$Y = X\beta^0 + Xb^0 + \varepsilon.$$

- We then consider the confounded model, i.e. a P_0 under which

$$Y = X\beta^0 + H\delta^0 + \nu.$$

- Throughout we consider the prior on $f = X\beta$ given by the model selection prior on β.

High-dimensional linear regression background

Compatibility

- For β^0 to be identifiable when $n < p$, need a notation of compatibility.
- If β^0 is sparse, then it is enough that something like the following compatibility number is bounded away from 0.
- For a matrix A, a model S, and a constant $M > 0$ we define the following compatibility number:

$$\phi_A(s) := \inf_{0 \neq |S_\beta| \leq s} \frac{\|A\beta\|_2}{\|A\| \|\beta\|_2},$$

where $\|A\| = \max_i \|A_{\cdot i}\|_2$.

- For $s-$sparse vectors

$$\|X(\beta_1 - \beta_2)\|_2 \geq \phi_X(s)\|X\| \|\beta_1 - \beta_2\|_2.$$

- We implicitly assume $\phi_X(s_0) > 0$ as it appears in the denominator of our rate, this is true with high probability for instance for orthogonal matrices and i.i.d. random matrices.

Prior conditions

- Recall the model selection prior which we place on β:

  1. The sparsity s of β is drawn according to some π_p a distribution on $\{1,\dots,p\}$. E.g. $\pi_p = \text{Binomial}(p, q)$, for some probability of inclusion q.
  2. The active set S given $|S| = s$ of β is drawn uniformly from the $\binom{p}{s}$ subsets of $\{1,\dots,p\}$ of size s.
  3. $\beta_i | S \overset{ind}{\sim} \begin{cases} \text{Lap}(\lambda), & i \in S, \\ \delta_0, & i \notin S\,. \end{cases}$

- Castillo et al. (2015) require that (λ, π_p) which determine the model selection prior satisfy:

$$\frac{\|X\|}{p} \leq \lambda/\sigma_\varepsilon \leq 2\bar{\lambda}, \qquad \bar{\lambda} = 2\|X\|\sqrt{\log p}\,.$$

and there exist constants $A_1, A_2, A_3, A_4 > 0$ that satisfy

$$A_1 p^{-A_3}\pi_p(s-1) \leq \pi_p(s) \leq A_2 p^{-A_4}\pi_p(s-1), \qquad s = 1,\dots,p\,.$$

- We make the same assumptions, but replace X in the first display with LX:

$$\frac{\|LX\|}{p} \leq \lambda/\sigma_\varepsilon \leq 2\bar{\lambda}, \qquad \bar{\lambda} = 2\|LX\|\sqrt{\log p}\,.$$

Castillo, I., Schmidt-Hieber, J., van der Vaart, A.. *Bayesian linear regression with sparse priors.* AOS, 2015.

High-dimensional linear regression

Theorem: Recovery in fixed design perturbed regression

Consider the perturbed linear model and suppose that the model selection prior assumptions hold.

Suppose further that the prior on b satisfies $\|LXb^0\|_2 = \mathcal{O}(1)$.

For $f = X\beta$, this is $\sqrt{n}\|f - f_0\|_{L,n,2}$ which we saw for GP regression

Then for sufficiently large $M > 0$

$$\sup_{\beta^0} E_0 \left[\Pi_\beta \left(\beta : \|LX(\beta - \beta^0)\|_2 > \frac{M\sigma_\varepsilon}{\phi_{LX}(|S_0|)} \sqrt{|S_0| \log p} \,|\, X, Y \right) \right] \to 0,$$

Note: constants have been simplified for discussion.

$$\sup_{\beta^0} E_0 \left[\Pi_\beta \left(\beta : \|\beta - \beta^0\|_2 > \frac{M\sigma_\varepsilon}{\phi^2_{LX}(|S_0|)} \frac{\sqrt{|S_0| \log p}}{\|LX\|} \,|\, X, Y \right) \right] \to 0.$$

Follows from $\|LX(\beta - \beta^0)\|_2 \geq \phi_{LX}(s)\|LX\|\|\beta - \beta^0\|_2$

IMPERIAL

High-dimensional linear regression

Theorem: Dimension in fixed design perturbed regression

Under the same assumptions, for some constant $M > 0$ which depends on $\phi_{LX}(s_0)$

$$\sup_{\beta^0} E_0 \left[\Pi_\beta \left(\beta : |S_\beta| > M |S_0| \, | X, Y \right) \right] \to 0.$$

High-dimensional linear regression

Corollary: Recovery in confounded regression

Consider the confounded model under our confounded model assumptions.

Suppose that $\sigma_{\max}(LX) = \mathcal{O}_p(\sqrt{p})$ and $\phi_{LX}(|S_0|)\|LX\| = \Omega_p(\sqrt{n\sigma_{\min}(\Sigma_X)})$.

Then for sufficiently large M and $\bar{\sigma}_\varepsilon^2 := \sigma_\nu^2 + \|\delta^0\|_2^2/\sigma_{\min}^2(\Gamma) < \infty$,

$$\sup_{\beta^0} E_0 \left[\Pi_\beta \left(\beta : \|\beta - \beta^0\|_2 > M \frac{\bar{\sigma}_\varepsilon}{\sqrt{\sigma_{\min}(\Sigma_X)}} \sqrt{\frac{|S_0| \log p}{n}} \,|\, X, Y \right) \right] \to 0.$$

- Obtain the usual rate (up to constants) of the Lasso estimator in the presence of no confounding.

Assumptions

Ćevid, D., Bühlmann, P., Meinhausen, N. *Spectral deconfounding via perturbed sparse linear models.* JMLR, 2020.

Assumptions

- The assumption $\phi_{LX}(|S_0|)\|LX\| = \Omega_p(\sqrt{n\sigma_{\min}(\Sigma_X)})$ looks complicated, but in fact is equivalent to

$$(\phi^*_{\tilde{\Sigma}})^2 = \Omega_p(\sigma_{\min}(\Sigma_X)),$$

given in Ćevid et al. (2020) and shown to be verified for the trim transform and another transform known as the puffer transform, which maps all singular values to 1.


Ćevid, D., Bühlmann, P., Meinhausen, N. *Spectral deconfounding via perturbed sparse linear models.* JMLR, 2020.