Iterative Proportional Scaling Revisited: A Modern Optimization Perspective

Yiyuan She, Shao Tang

Department of Statistics, Florida State University

What is IPS?

- We would like to adjust a given table $\boldsymbol{q}$ (say, a table of ones) get a new table $\boldsymbol{\mu}$ satisfying a set of given margins
- Consider a three-way example. Let $\{m_{ij+}, m_{+jk}, m_{i+k}\}$ be the pre-specified margins, possibly calculated from $[m_{ijk}]$
- Starting with $\boldsymbol{\mu}^{(0)} = \boldsymbol{q}$ (often $\mathbf{1}$), IPS performs the following **scaling** operations iteratively for any $t \geq 0$:

$$\mu_{ijk}^{(3t+1)} = \mu_{ijk}^{(3t)} \frac{m_{ij+}}{\mu_{ij+}^{(3t)}}, \quad \forall (i,j,k)$$

$$\mu_{ijk}^{(3t+2)} = \mu_{ijk}^{(3t+1)} \frac{m_{i+k}}{\mu_{i+k}^{(3t+1)}}, \quad \forall (i,j,k)$$

$$\mu_{ijk}^{(3t+3)} = \mu_{ijk}^{(3t+2)} \frac{m_{+jk}}{\mu_{+jk}^{(3t+2)}}, \quad \forall (i,j,k)$$

Who Cares?

- IPS has widespread applications in computer science (Rote & Zachariasen 07), econometrics (Lahr & Mesnarad 04), and mathematics (Sinkhorn & Knopp 67, Pretzel 80)
- Matrix raking
- Contingency table analysis: test of association
- Log-linear model fitting: estimation and goodness-of-fit
  - X^2, G^2: expected frequencies. Log-linear: $\boldsymbol{X}^T\boldsymbol{\mu} = \boldsymbol{X}^T\boldsymbol{y}$

History

- Deming & Stephan (40): table adjustment (aiming at X^2)
- Ireland & Kullback (68): **KL** entropy subject to constraints
- Bishop et al (75): log-linear **likelihood** for tables
- Csiszar (75): duality holds though Lagrangian cannot apply
- Convergence studies undergo a long history
  - Brown (59), Birch (63), Fienberg (70), Haberman (74), Bishop et al (75), Csiszar (75), Ruschendorf (95), Pukelsheim & Simeone (09), Kurras (15), just to name a few
- **Extensions**:
  - Generalized Iterative Scaling (**GIS**) (Darroch & Ratcliff 72). Improved Iterative Scaling (**IIS**) (Pietra et al 97)

Big Data Era

- Nowadays, IPS is not as popular as **Newton**-type algorithms, and has become an outdated method

  "*primarily of historical interest*" (Agresti 12)

- The algorithm is pretty motivating in the **big** data era
- For a 5-way table with 10 levels each, a model including up to three-term interactions has 8,146 parameters
- Hessian: more than 10^7 entries..
- First-order methods? No universal step size.

Pros and Cons of (Plain) IPS

Pros

- No step-size parameter
- Memory saving
- Good numerical stability

Cons

- No coefficient estimate (or standard errors) but the mean
- Narrow scope: only applies to tables (MLE, no shrinkage)
- Slow convergence, though cost-effective per iteration

We would like to study IPS from a modern optimization perspective to improve and generalize this elegant algorithm

Setting

- Given a multi-way table, we can introduce dummy variables to form a binary design matrix $\boldsymbol{X} \in \mathbb{R}^{N \times p}$.
- Accordingly, we vectorize the table of counts to $\boldsymbol{n} \in \mathbb{R}^N$.
- IPS (a general version) can be described by
  1. $\log \boldsymbol{\mu}^{(0)} \in \log \boldsymbol{q} + \mathcal{R}(\boldsymbol{X})$ and $t \leftarrow 1$.
  2. With $\boldsymbol{\mu}^{(t,0)} \leftarrow \boldsymbol{\mu}^{(t-1)}$, $\boldsymbol{\mu}^{(t,j)} = [\mu_i^{(t,j)}]$ $(1 \le j \le p)$ are obtained sequentially below:

$$\mu_i^{(t,j)} = \begin{cases} \mu_i^{(t,j-1)} \dfrac{\langle \boldsymbol{x}_j, \boldsymbol{n} \rangle}{\langle \boldsymbol{x}_j, \boldsymbol{\mu}^{(t,j-1)} \rangle} & x_{ij} = 1 \\ \mu_i^{(t,j-1)} & x_{ij} = 0, \end{cases} \quad 1 \le i \le N,$$

  3. $\boldsymbol{\mu}^{(t+1)} \leftarrow \boldsymbol{\mu}^{(t,p)}$. If not converged, $t \leftarrow t+1$, go to Step 2.

Log-Affine Models

- Assume n_i are independent and follow Poisson distributions

$$n_i \sim Poi(\mu_i), \text{ with } \log \boldsymbol{\mu} - \log \boldsymbol{q} = \boldsymbol{X}\boldsymbol{\beta} \text{ for some } \boldsymbol{\beta} \in \mathbb{R}^p$$

- When $\boldsymbol{q} = \mathbf{1}$, it degenerates to a log-linear model
- The maximum likelihood problem for $\boldsymbol{\mu}$ is then

$$\begin{aligned} &\min_{\boldsymbol{\mu}} l(\boldsymbol{\mu}) \triangleq -\langle \boldsymbol{n}, \log \boldsymbol{\mu} \rangle + \langle \mathbf{1}, \boldsymbol{\mu} \rangle \\ &\text{s.t. } \boldsymbol{\mu} \in \mathcal{M} \triangleq \{\boldsymbol{\mu} \,|\, \log \boldsymbol{\mu} \in \log \boldsymbol{q} + \mathcal{R}_{\boldsymbol{X}}, \boldsymbol{\mu} \succeq \mathbf{0}\}. \end{aligned}$$

- Table-derived designs are always binary. In general, we could have binary/nonnegative/general designs

Formulation through **Coefficients**

- Seen from the initialization requirement and the model description, it is very natural to define the problem via $\boldsymbol{\beta}$

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} l(\boldsymbol{\beta}) = -\langle \boldsymbol{n}, \boldsymbol{X}\boldsymbol{\beta}\rangle + \langle \boldsymbol{q}, \exp(\boldsymbol{X}\boldsymbol{\beta})\rangle$$

- A simple cyclic coordinate descent algorithm

$$\beta_j^{(t+1)} \in \arg\min_{\beta_j} l(\beta_1^{(t+1)}, \ldots, \beta_{j-1}^{(t+1)}, \beta_j, \beta_{j+1}^{(t)}, \ldots, \beta_p^{(t)})$$

- Initialize through $\boldsymbol{\beta}$; then set $\boldsymbol{\mu}^{(0)} \leftarrow \boldsymbol{q} \circ \exp(\boldsymbol{X}\boldsymbol{\beta}^{(0)})$
- Mean-update: $\boldsymbol{\mu}^{(t,j)} \leftarrow \boldsymbol{\mu}^{(t,j-1)} \circ \exp[\boldsymbol{x}_j(\beta_j^{(t+1)} - \beta_j^{(t)})]$ (scaling)

- $\partial l/\partial \beta_j = 0$ gives

$$(\boldsymbol{x}_j \circ \boldsymbol{\mu}^{(t,j-1)})^T \exp\{\boldsymbol{x}_j(\beta_j - \beta_j^{(t)})\} - \boldsymbol{x}_j^T \boldsymbol{n} = 0.$$

- With $x_{ij} = 0$ or 1, we have $x_{ij} \exp(x_{ij}\beta_j) = x_{ij} \exp(\beta_j)$ or $\beta_j^{(t+1)} = \beta_j^{(t)} + \log(\frac{\boldsymbol{x}_j^T \boldsymbol{n}}{\boldsymbol{x}_j^T \boldsymbol{\mu}^{(t,j-1)}})$, and so

$$\boldsymbol{\mu}^{(t,j)} = \boldsymbol{\mu}^{(t,j-1)} \circ \exp\{\boldsymbol{x}_j \log[(\boldsymbol{x}_j^T \boldsymbol{n})/(\boldsymbol{x}_j^T \boldsymbol{\mu}^{(t,j-1)})]\}$$

  which is exactly IPS!
- We get a novel CD characterization of IPS

Benefits

- Flexibility in initialization: an arbitrary $\boldsymbol{\beta}^{(0)} \in \mathbb{R}^p$ suffices
- IPS can produce **parameter** estimates (as well as an estimate of the asymptotic covariance)!
- No need to scan the data repeatedly to fulfill all margins Multiple components of $\boldsymbol{\mu}$ can be simultaneously updated
- The CD perspective also makes it easy to generalize and improve IPS as well as facilitating theoretical investigations

IPS for Minimizing X^2

- Deming & Stephan (40) conjectured that IPS minimizes the X^2 statistic, later known to be incorrect
- Applying CD to $\min_{\boldsymbol{\beta}\in\mathbb{R}^p} X^2(\boldsymbol{\beta}) \triangleq \sum_i \frac{[n_i-\mu_i(\boldsymbol{\beta})]^2}{\mu_i(\boldsymbol{\beta})}$ gives

$$\boldsymbol{\mu}^{(t,j)} = \boldsymbol{\mu}^{(t,j-1)} \circ \exp\{\frac{\boldsymbol{x}_j}{2} \log[\boldsymbol{x}_j^T(\boldsymbol{n} \circ \boldsymbol{n} \oslash \boldsymbol{\mu}^{(t,j-1)})/(\boldsymbol{x}_j^T \boldsymbol{\mu}^{(t,j-1)})]\}$$

- This is the desired X^2-version of IPS

Theoretical Properties

- From our derivation, it is easy to show for $\forall t \geq 0$, $j \in [p]$

$$l(\boldsymbol{\beta}^{(t,j)}) \geq l(\boldsymbol{\beta}^{(t+1,j)})$$

- With the intercept present, we can further show that the same conclusion holds for $G^2 \triangleq 2\sum_i n_i \log(n_i/\mu_i)$
  - Many researchers observed that G^2 decreases during IPS
  - This is just a natural outcome of its CD nature
- Furthermore, the sequence of iterates converges to the unique optimal solution, at least **linearly**

Accelerations

- Speeding the convergence of IPS is crucial on big tables
- **Cyclic** update has slow (worst-case) convergence rate
- One idea: pick the coordinate with the "largest" derivative

$$\text{Gauss-Southwell rule: } j = \arg\max_j |\nabla_j l(\boldsymbol{\beta}^{(t)})|$$

- But this is not worthy since when the full gradient is available, one can directly update the whole $\boldsymbol{\beta}$-vector
- A powerful idea is to apply randomization!
- In theory, randomized coordinate descent can avoid the worst-case bounds (Nesterov 12). We call it **A-IPS**

An MM Viewpoint

- Majorization-Minimization algorithms are popular alternatives to EM algorithms
- Rather than directly minimizing $l(\boldsymbol{\beta})$, we construct a surrogate function $g(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{-})$ satisfying

$$g(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{-}) \geq l(\boldsymbol{\beta}) \ , \ g(\boldsymbol{\beta} \mid \boldsymbol{\beta}) = l(\boldsymbol{\beta})$$

- Let $\boldsymbol{\beta}^{(t+1)} \in \arg\min_{\boldsymbol{\beta}} g(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(t)})$. Then the function value is guaranteed to be decreasing
- Once getting $\boldsymbol{\beta}^{(t+1)}$, update $\boldsymbol{\mu}$ via **proportional scaling**

$$\boldsymbol{\mu}^{(t+1)} = \boldsymbol{\mu}^{(t)} \circ \exp[\sum_{j} \boldsymbol{x}_j(\beta_j^{(t+1)} - \beta_j^{(t)})]$$

- The problem boils down to the design of proper surrogates

- Recall the objective function: $l(\boldsymbol{\beta}) = \sum_i l_i$ with $l_i = q_i \exp(\tilde{\boldsymbol{x}}_i^T \boldsymbol{\beta}) - n_i \tilde{\boldsymbol{x}}_i^T \boldsymbol{\beta}$.
- Applying Jensen's inequality gives

$$
\begin{aligned}
l_i &\leq -n_i \tilde{\boldsymbol{x}}_i^T \boldsymbol{\beta} + \sum_j \alpha_{ij} q_i \exp[\frac{x_{ij}}{\alpha_{ij}}(\beta_j - \beta_j^-) + \tilde{\boldsymbol{x}}_i^T \boldsymbol{\beta}^-] \\
&= -n_i \tilde{\boldsymbol{x}}_i^T \boldsymbol{\beta} + \sum_j \alpha_{ij} \mu_i^- \exp[\frac{x_{ij}}{\alpha_{ij}}(\beta_j - \beta_j^-)],
\end{aligned}
$$

where α_{ij} satisfy $\alpha_{ij} \geq 0$, $\sum_{j=1}^p \alpha_{ij} = 1$ for all i.

First Try

- Let $\alpha_{ij} = 1/p$. Then for all binary designs, we get

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + \frac{1}{p} \log[(\boldsymbol{X}^T \boldsymbol{n}) \oslash (\boldsymbol{X}^T \boldsymbol{\mu}^{(t)})]$$

$$\boldsymbol{\mu}^{(t+1)} = \boldsymbol{\mu}^{(t)} \circ (\exp\{\frac{1}{p} \boldsymbol{X} \log[(\boldsymbol{X}^T \boldsymbol{n}) \oslash (\boldsymbol{X}^T \boldsymbol{\mu}^{(t)})]\})$$

- This MM algorithm resembles IPS, but updates all components synchronously with a smaller stepsize $(1/p)$

Non-negative Designs

- A perhaps better choice is to set $\alpha_{ij} = x_{ij}/x_{i+}$, leading to

$$l_i \leq -\sum_j n_i x_{ij} \beta_j + \sum_j \frac{x_{ij}}{x_{i+}} \mu_i^- \exp[x_{i+}(\beta_j - \beta_j^-)]$$

- However, the g-optimization has no a closed-form solution
- We can majorize $\exp[x_{i+}(\beta_j - \beta_j^-)]/x_{i+}$ further. For any a, b with $0 < a \leq b$, $\exp(at)/a - 1/a \leq \exp(bt)/b - 1/b$
- With $a = x_{i+}$ and $b = \max_i x_{i,+} = \|\boldsymbol{X}\|_\infty \triangleq R$, we get

$$g_2(\boldsymbol{\beta} \mid \boldsymbol{\beta}^-) = -\boldsymbol{n}^T \boldsymbol{X} \boldsymbol{\beta} + \sum_{i,j} x_{ij} \mu_i^- \left\{ \frac{\exp[R(\beta_j - \beta_j^-)]}{R} - \frac{1}{R} + \frac{1}{x_{i+}} \right\}$$

which has an explicit solution for any **nonnegative** $\boldsymbol{X}$!

Extending the **Generalized Iterative Scaling**

- The iterates are then given by

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + \frac{1}{R}\log[(\boldsymbol{X}^T\boldsymbol{n}) \oslash (\boldsymbol{X}^T\boldsymbol{\mu}^{(t)})]$$

$$\boldsymbol{\mu}^{(t+1)} = \boldsymbol{\mu}^{(t)} \circ (\exp\{\frac{1}{R}\boldsymbol{X}\log[(\boldsymbol{X}^T\boldsymbol{n}) \oslash (\boldsymbol{X}^T\boldsymbol{\mu}^{(t)})]\})$$

- Larger stepsize than the previous one
- This takes the celebrated **GIS** (Darroch & Ratcliff 72) as a particular instance
- In fact, using the MM trick, we can even extend IPS to fit any log-affine models with arbitrary designs

A **Reparametrization** Trick

- Assume the intercept is present: $\boldsymbol{X} = [\mathbf{1}\ \mathring{\boldsymbol{X}}]$, $\boldsymbol{\beta} = [\beta_0\ \mathring{\boldsymbol{\beta}}^T]^T$
- Introducing $\alpha \triangleq \beta_0 + \log\langle \boldsymbol{q}, \exp(\mathring{\boldsymbol{X}}\mathring{\boldsymbol{\beta}})\rangle$, or $\exp(\alpha) = \langle \boldsymbol{q}, \exp(\beta_0\mathbf{1} + \mathring{\boldsymbol{X}}\mathring{\boldsymbol{\beta}})\rangle = \langle \mathbf{1}, \boldsymbol{\mu}\rangle$, $l(\boldsymbol{\beta})$ becomes

$$-\langle \boldsymbol{n}, \mathring{\boldsymbol{X}}\mathring{\boldsymbol{\beta}}\rangle + \langle \mathbf{1}, \boldsymbol{n}\rangle \log\langle \boldsymbol{q}, \exp(\mathring{\boldsymbol{X}}\mathring{\boldsymbol{\beta}})\rangle + \exp(\alpha) - \langle \mathbf{1}, \boldsymbol{n}\rangle\alpha$$

- It suffices to study

$$\min_{\mathring{\boldsymbol{\beta}}} L(\mathring{\boldsymbol{\beta}}) \triangleq -\langle \boldsymbol{n}, \mathring{\boldsymbol{X}}\mathring{\boldsymbol{\beta}}\rangle + \langle \mathbf{1}, \boldsymbol{n}\rangle \log\langle \boldsymbol{q}, \exp(\mathring{\boldsymbol{X}}\mathring{\boldsymbol{\beta}})\rangle$$

- This gives a really *nice* form in optimization!

- Due to the **concavity** of the log function, for any $\zeta > 0$, we have $\log(\zeta x) \leq \zeta x - 1$ (with equality at $\zeta = 1/x$)
- $L(\mathring{\boldsymbol{\beta}}) \leq -\langle \mathbf{1}, \boldsymbol{n} \rangle \log \zeta - \langle \boldsymbol{n}, \mathring{\boldsymbol{X}} \mathring{\boldsymbol{\beta}} \rangle + \langle \mathbf{1}, \boldsymbol{n} \rangle [\zeta \langle \boldsymbol{q}, \exp(\mathring{\boldsymbol{X}} \mathring{\boldsymbol{\beta}}) \rangle - 1]$
- Assuming the non-negative setting and applying Jensen's inequality with $\alpha_{ij} = \mathring{x}_{ij} / \mathring{x}_{i+}$, we get

$$g_3(\mathring{\boldsymbol{\beta}} \mid \mathring{\boldsymbol{\beta}}^-) = -\langle \boldsymbol{n}, \mathring{\boldsymbol{X}} \mathring{\boldsymbol{\beta}} \rangle - \langle \mathbf{1}, \boldsymbol{n} \rangle (\log \zeta + 1) + \zeta \langle \mathbf{1}, \boldsymbol{n} \rangle \sum_{i \in [N], j \in [p-1]} \frac{\mathring{x}_{ij}}{\mathring{x}_{i+}} \mathring{\mu}_i^- \exp[\mathring{x}_{i+} (\mathring{\beta}_j - \mathring{\beta}_j^-)]$$

- The only choice of ζ to guarantee that g_3 is a surrogate is $\zeta = 1 / \langle \boldsymbol{q}, \exp(\mathring{\boldsymbol{X}} \mathring{\boldsymbol{\beta}}^-) \rangle$

Recovering the **Improved Iterative Scaling**

- $\mathring{\beta}_j^{(t+1)}$ is determined by

$$\frac{\exp(\beta_0^{(t)})\langle \mathbf{1}, \boldsymbol{n}\rangle}{\langle \mathbf{1}, \boldsymbol{\mu}^{(t)}\rangle} \sum_i \mathring{x}_{ij} \mathring{\mu}_i^{(t)} \exp[\mathring{x}_{i+}(\mathring{\beta}_j^{(t+1)} - \mathring{\beta}_j^{(t)})] = \sum_i n_i \mathring{x}_{ij}$$

- Interestingly, this MM algorithm fully restores the **IIS** (Pietra et al 97), if we add a mean-normalization and an intercept update
- These two steps are however unnecessary in our algorithm

Quadratic Surrogates

- The re-parameterized form allows for linearization

$$Q(\mathring{\boldsymbol{\beta}} \mid \mathring{\boldsymbol{\beta}}^{-}) = L(\mathring{\boldsymbol{\beta}}^{-}) + \langle \mathring{\boldsymbol{\beta}} - \mathring{\boldsymbol{\beta}}^{-}, \nabla_{\mathring{\boldsymbol{\beta}}} L(\mathring{\boldsymbol{\beta}}^{-}) \rangle + \frac{1}{2} (\mathring{\boldsymbol{\beta}} - \mathring{\boldsymbol{\beta}}^{-})^T \boldsymbol{W} (\mathring{\boldsymbol{\beta}} - \mathring{\boldsymbol{\beta}}^{-})$$

- Q is a valid surrogate function provided that

$$\langle \mathbf{1}, \boldsymbol{n} \rangle \mathring{\boldsymbol{X}}^T \left[\frac{\operatorname{diag}(\mathring{\boldsymbol{\mu}})}{\langle \mathbf{1}, \mathring{\boldsymbol{\mu}} \rangle} - \frac{\mathring{\boldsymbol{\mu}} \mathring{\boldsymbol{\mu}}^T}{\langle \mathbf{1}, \mathring{\boldsymbol{\mu}} \rangle^2} \right] \mathring{\boldsymbol{X}} \preceq \boldsymbol{W} \ , \ \forall \mathring{\boldsymbol{\mu}} \succeq \mathbf{0}$$

- An obvious choice is $\boldsymbol{W} = (\langle \mathbf{1}, \boldsymbol{n} \rangle \| \mathring{\boldsymbol{X}} \|_2^2 / 2) \boldsymbol{I}$
- A finer one:

$$\boldsymbol{W} = \mathring{\boldsymbol{X}}^T (\langle \mathbf{1}, \boldsymbol{n} \rangle \boldsymbol{I} - \mathbf{1}\mathbf{1}^T) \mathring{\boldsymbol{X}} / 2$$

Q-IPS

- Fancier techniques like Nesterov's **second acceleration** can be applied to speed the convergence
- Initialization: $t \leftarrow 0$, $\mathring{\boldsymbol{\beta}} \leftarrow \mathring{\boldsymbol{\beta}}^{(0)}$, $\mathring{\boldsymbol{\eta}}^{(0)} \leftarrow \mathring{\boldsymbol{\beta}}^{(0)}$, $\theta_0 = 1$
- The momentum trick to guarantee faster convergence:
  1. $\mathring{\boldsymbol{\alpha}}^{(t)} \leftarrow (1-\theta_t)\mathring{\boldsymbol{\beta}}^{(t)} + \theta_t \mathring{\boldsymbol{\eta}}^{(t)}$
  2. $\mathring{\boldsymbol{\eta}}^{(t+1)} \leftarrow$
     $\mathring{\boldsymbol{\eta}}^{(t)} - \theta_t^{-1}\boldsymbol{W}^{-1}\{-\mathring{\boldsymbol{X}}^T\boldsymbol{n} + \frac{\langle \mathbf{1}, \boldsymbol{n}\rangle}{\langle \boldsymbol{q}, \exp(\mathring{\boldsymbol{X}}\mathring{\boldsymbol{\alpha}}^{(t)})\rangle}\mathring{\boldsymbol{X}}^T[\boldsymbol{q} \circ \exp(\mathring{\boldsymbol{X}}\mathring{\boldsymbol{\alpha}}^{(t)})]\}$
  3. $\mathring{\boldsymbol{\beta}}^{(t+1)} \leftarrow (1-\theta_t)\mathring{\boldsymbol{\beta}}^{(t)} + \theta_t \mathring{\boldsymbol{\eta}}^{(t+1)}$
  4. $\mathring{\boldsymbol{\mu}}^{(t+1)} \leftarrow \mathring{\boldsymbol{\mu}}^{(t)} \circ \exp[\mathring{\boldsymbol{X}}(\mathring{\boldsymbol{\beta}}^{(t+1)} - \mathring{\boldsymbol{\beta}}^{(t)})]$
  5. $\theta_{t+1} \leftarrow (\sqrt{\theta_t^4 + 4\theta_t^2} - \theta_t^2)/2$
  6. $t \leftarrow t+1$

Divide-and-Conquer for Big Data

- The CD idea can be extended further to BCD
- We recommend **randomized** BCD. But the key is not to pick a block at random, but to perform random blocking
- This **B**-IPS algorithm can combine the virtues of IPS and Newton-like algorithms and is extremely scalable
  - Typically we adopt block sizes ≈ 200

Shrinkage Estimation

- We use the ℓ_1-penalized contingency **table** analysis as an example (the ℓ_2-penalty is much easier to handle)
- To identify significant association terms, we can solve

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} l_1(\boldsymbol{\beta}) \triangleq -\langle \boldsymbol{n}, \boldsymbol{X}\boldsymbol{\beta}\rangle + \langle \boldsymbol{q}, \exp(\boldsymbol{X}\boldsymbol{\beta})\rangle + \sum_{j=1}^{p} \lambda_j |\beta_j|,$$

  where the penalty is not imposed on the intercept ($\lambda_1 = 0$)
- IPS/CD: $\min_{\beta_j\in\mathbb{R}} l_1(\beta_0^-, \ldots, \beta_{j-1}^-, \beta_j, \beta_{j+1}^- \ldots, \beta_{p-1}^-)$

Poisson 'Soft-Thresholding' Operator

- Define

$$\Theta(\boldsymbol{x}, \boldsymbol{n}, \boldsymbol{q}; \lambda) = \begin{cases} \log \frac{\langle \boldsymbol{x}, \boldsymbol{n} \rangle - \lambda \operatorname{sgn} \langle \boldsymbol{x}, \boldsymbol{n} - \boldsymbol{q} \rangle}{\langle \boldsymbol{x}, \boldsymbol{q} \rangle}, & |\langle \boldsymbol{x}, \boldsymbol{n} - \boldsymbol{q} \rangle| \geq \lambda \\ 0, & |\langle \boldsymbol{x}, \boldsymbol{n} - \boldsymbol{q} \rangle| < \lambda \end{cases}$$

- Then, an ℓ_1-IPS algorithm for tables can be developed (similar to the fast CD lasso algorithm)

$$\beta_j = \Theta(\boldsymbol{x}, \boldsymbol{n}, \boldsymbol{q} \circ \exp(\boldsymbol{X}_{\setminus j} \boldsymbol{\beta}_{\setminus j}^{-}); \lambda), \ j = 1, \ldots, p$$

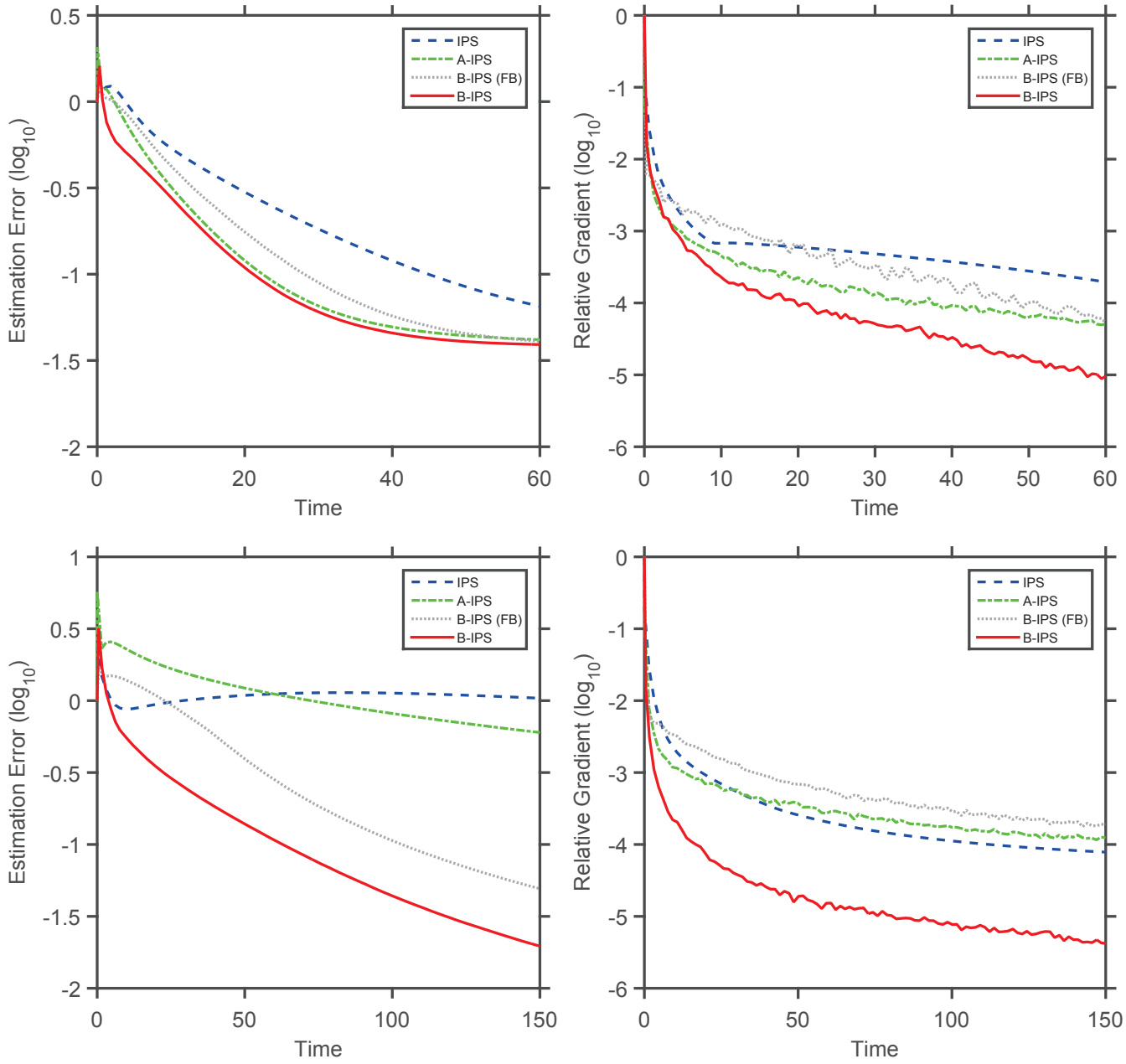

Figure: Performance comparison in two (small) table settings ($p = 523$, $N = 100K$)

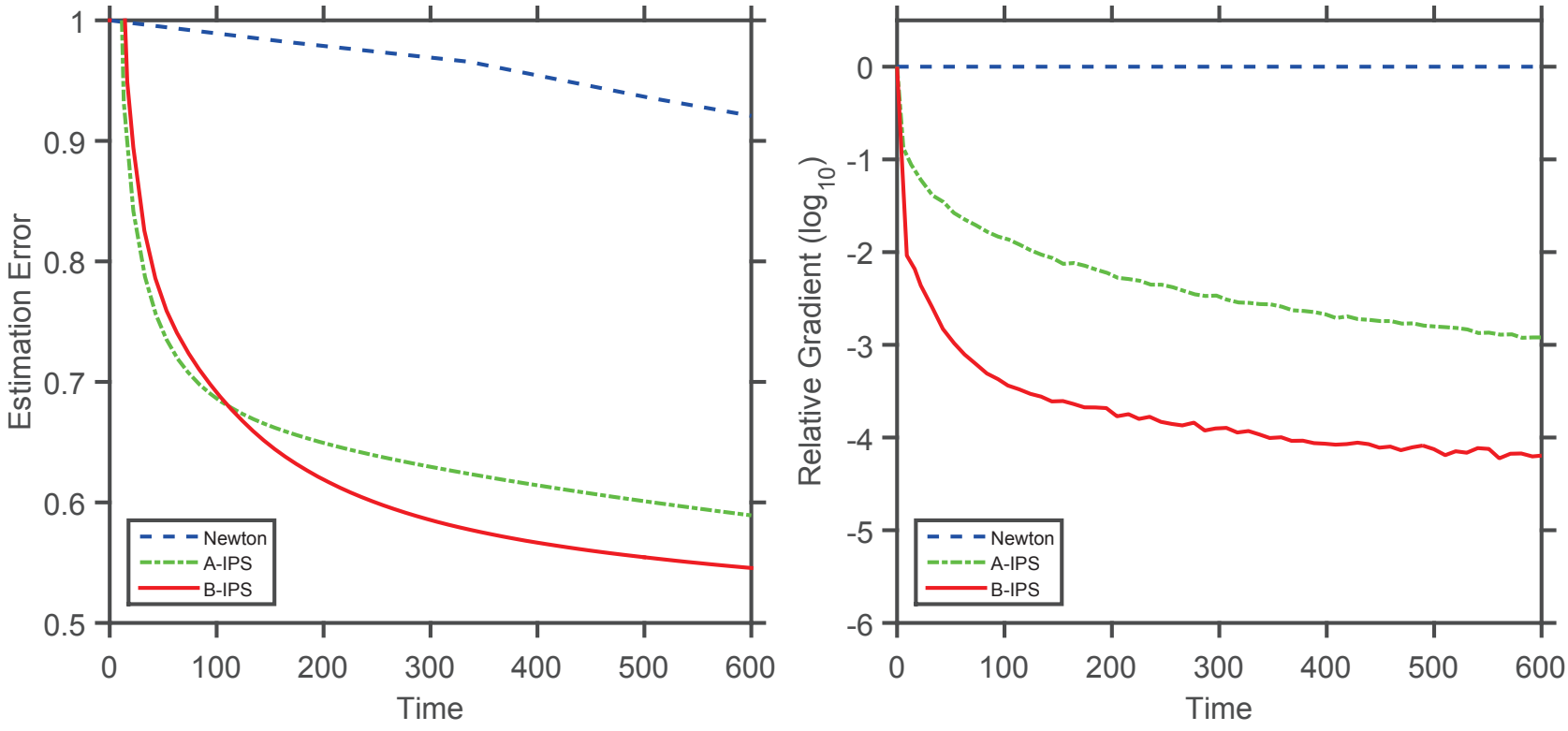

Figure: Large table setting ($p = 8,146,\ N = 100K$).

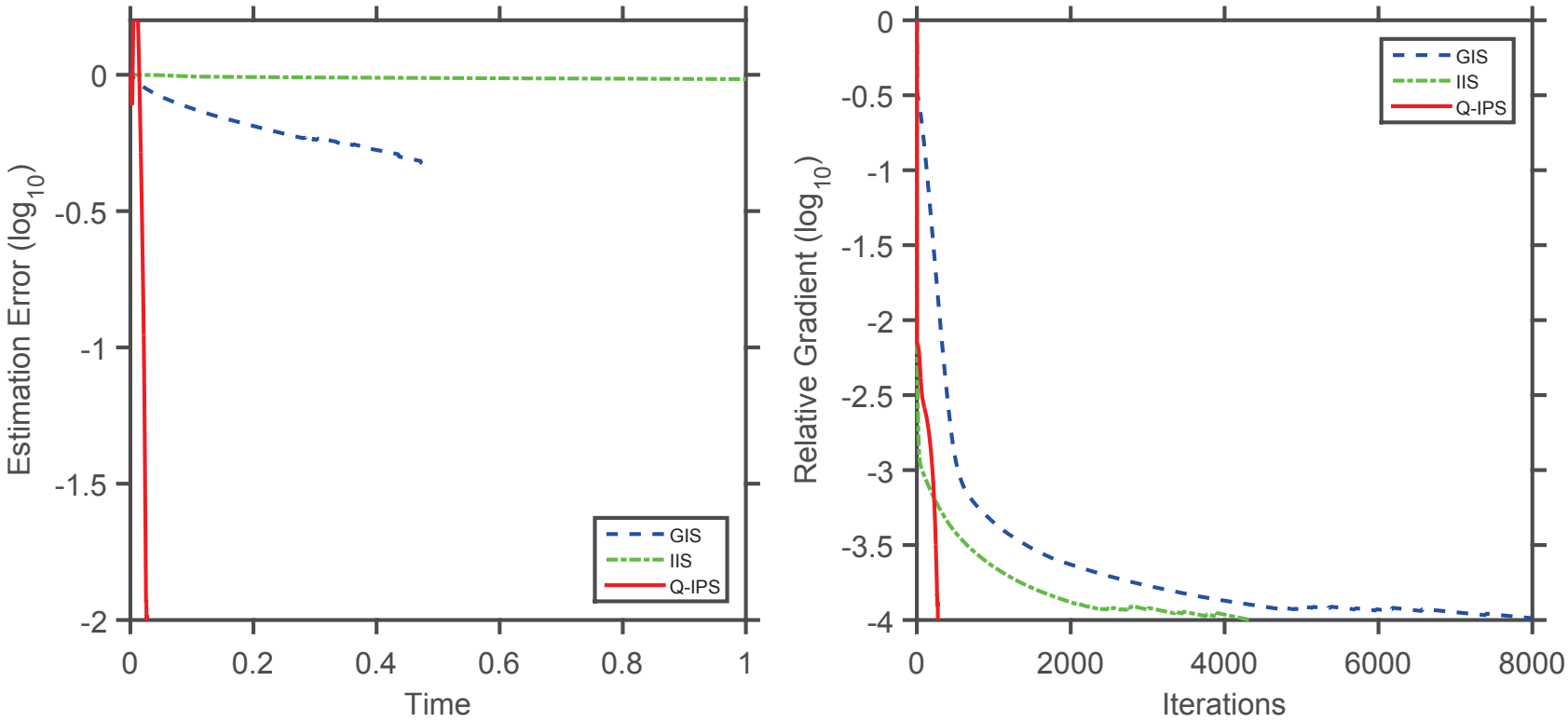

Figure: Nonnegative design (small): $N = 1000, p = 100$.

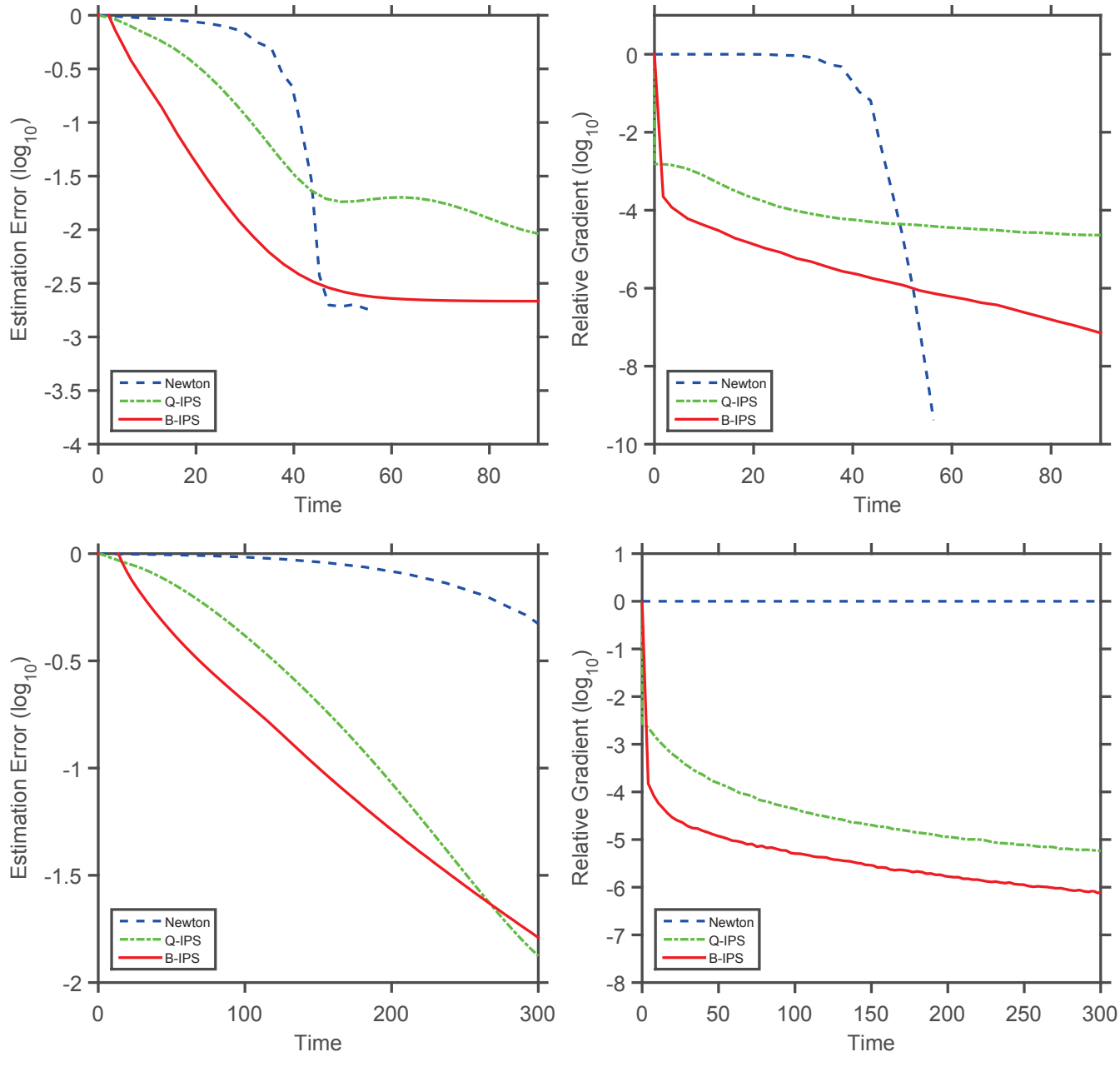

Figure: Large non-negative designs ($p = 2,000$, $p = 4,000$)

- General designs

	$p = 1,000$ tol = 1e-6		$p = 4,000$ tol = 1e-6		$p = 10,000$ $T_{\max} = 600$	
	Time	estErr	Time	estErr	relGrad	estErr
Newton	35.2	0.12	528.5	0.1	–	–
L-BFGS	35.6	2.3	276.5	2.6	7622	$9.3e+2$
B-IPS	16.8	3.2	116.5	23.2	5.70	$1.0e+2$

ℓ_1 Solution Path on Bank Marketing Data

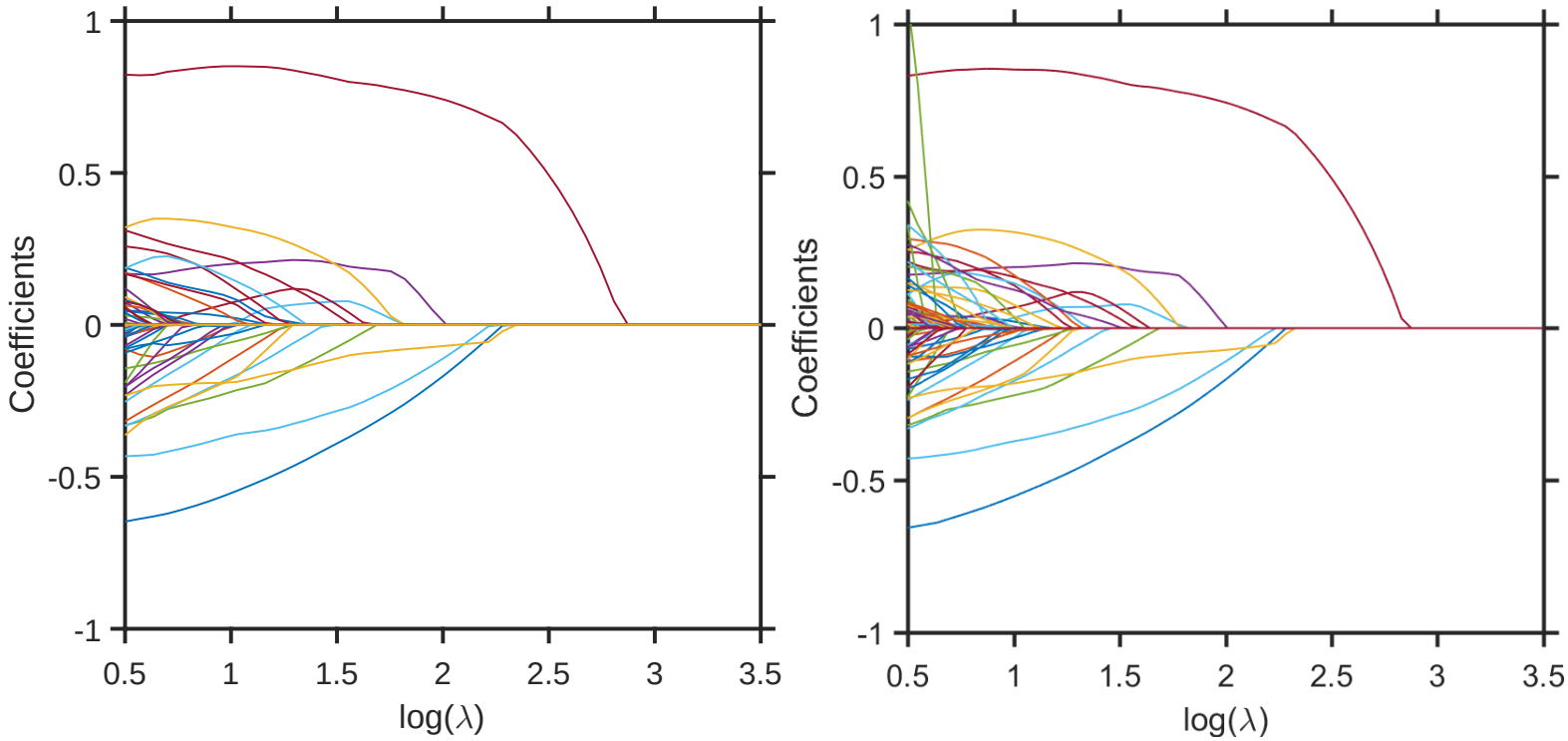

Figure: ℓ_1 solution path for two-way ($p = 739$) and three-way ($p = 5,784$) homogeneous association models on the left and right panel, respectively.

ℓ_1 Selected Variables

Main	month = mar (+) month = may (−) month = oct (+) month = nov (−) education = univ (+) poutcome = success (+) default = unkown (−) contact = telephone (−) loan = yes (−)
Two-way	(poutcome = non-exist) ∗ month = mar (+) (poutcome = non-exist) ∗ month = oct (+) (poutcome = non-exist) ∗ education = univ (+) (poutcome = non-exist) ∗ contact = telephone (−) (poutcome = non-exist) ∗ housing = unkown (−)
Three-way	(job = technician) ∗ (education = professional) ∗ (marital = married) (+) (poutcome = non-exist) ∗ education = univ ∗ (marital = single) (+)

Table: Selected variables with signs of coefficients in parentheses.

Bank Marketing Data Analysis

- `month` plays an important role in the marketing campaign.
- Contact communication type being cellular and the outcome of previous marketing campaign being successful are favorable factors for successful subscription.
- All selected two-way interactions involve `poutcome` = `non-exist`.
- The three-way terms selected show no collinearity with the other selected variables.

Summary

- We revisited the conventional IPS from a modern optimization perspective
- Based on a coordinate descent characterization, we showed that IPS can deliver estimates of coefficients
- From a majorization-minimization standpoint, IPS can be substantially generalized to handle any log-affine models
- IPS can be accelerated in various ways to provide extremely scalable algorithms in big data applications
- Our techniques apply to shrinkage estimation on count data