

Penalized Least Squares Methods for Latent Variables Models

**A discussion of two papers
by S. Schennach and V. Chernozhukov**

Stéphane Bonhomme
CEMFI

World Congress of the Econometric Society
Shanghai, August 2010

Schennach: latent variables models (LVM)

- Derives several nonparametric identification results. As an important example: Hu and Schennach (08)'s very general results (under conditional independence restrictions).
- Focus on measurement error models. However, the proposed ideas also apply to group and panel data (with fixed effects), dynamic decision models (unobserved states: Hu and Shum 10)...
- Estimation is not straightforward:
 - Characteristic-function based estimators in additive models.
 - Nonlinear models: sieve Maximum Likelihood is an option (Shen 97, Chen 07). However: 1) the estimation problem is nonlinear; and 2) the dimensionality of the problem is high.

Chernozhukov: ℓ^1 penalized least squares (Lasso)

- Focus on linear regression models with potentially many regressors.
- Lasso idea (Tibshirani 96): add the ℓ^1 norm of the parameter (i.e., $\sum_j |\beta_j|$) to the Least Squares criterion. This is easy to minimize (convex programming), and selects corner solutions (many $\hat{\beta}_j$'s are typically zeros).
- Belloni and Chernozhukov (10a, 10b) apply the Lasso idea to quantile regression, and introduce the post-Lasso (to correct for bias).
- Economic applications: “robust” determinants (e.g., growth regression), nonlinear regression (approximating the regression function in a dictionary), many instruments problem...

Combining the two approaches

- LVM are high-dimensional, due to the presence of unknown distribution functions.

⇒ Why not using the Lasso to estimate those models?

- Outline of the presentation:
 1. A simple latent variable model
 2. Penalized Least Squares density estimation
 3. Open issues for application to LVM

Measurement error

- Suppose that we are interested in documenting the relationship between an outcome Y and a covariate X^* .
- We do not observe X^* , but only an imperfect measure X .
- Assumptions:
 1. Y is independent of X given X^* .
 2. The distribution function of X given X^* is known (or estimated).

For example: auxiliary data, or repeated measures of X^* , are available.

Penalized likelihood

- A popular approach is to assume that the distributions belong to “rich” families depending on some parameter γ , and to estimate:

$$\hat{\gamma} = \operatorname{argmax}_{\gamma \in \Gamma} \sum_{i=1}^N \log \int \underbrace{f(Y_i | X^*; \gamma)}_{=f(Y_i | X_i, X^*; \gamma)} f(X^*; \gamma) \underbrace{f(X_i | X^*)}_{\text{known}} dX^*.$$

- Various choices of Γ yield:
 - Parametric likelihood (Γ fixed), sieve ML (Γ expands with N).
 - Ridge ($\Gamma = \{\gamma, \sum_j |\gamma_j|^2 \leq t_N\}$), or Lasso ($\Gamma = \{\gamma, \sum_j |\gamma_j| \leq t_N\}$).
- Nonlinear problem \Rightarrow computationally challenging.

A linear model

- Let $\{\psi_k(Y, X^*)\}$ denote a rich set of functions, possibly non-orthogonal. Specify:

$$f(Y, X^*) = \sum_k a_k \psi_k(Y, X^*).$$

- The distribution function of (Y, X) is linear in $\{a_k\}$:

$$\begin{aligned} f(Y, X) &= \int f(Y|X, X^*) f(X^*) f(X|X^*) dX^* \\ &= \int f(Y, X^*) f(X|X^*) dX^* \\ &= \sum_k a_k \underbrace{\int \psi_k(Y, X^*) f(X|X^*) dX^*}_{\equiv \varphi_k(Y, X), \text{ known}}. \end{aligned}$$

(Note: rescale φ_k such that $\sup |\varphi_k| = 1$)

- Problem: how to exploit linearity in estimation?

Minimization of the ℓ^2 distance

- Note that $\{a_k\}$ minimizes the squared ℓ^2 distance:

$$\begin{aligned} \left\| f(Y, X) - \sum_k a_k \varphi_k(Y, X) \right\|_2^2 &= \iint \left(f(Y, X) - \sum_k a_k \varphi_k(Y, X) \right)^2 dY dX \\ &= \|f(Y, X)\|_2^2 + \left\| \sum_k a_k \varphi_k(Y, X) \right\|_2^2 - 2\mathbb{E} \left(\sum_k a_k \varphi_k(Y, X) \right). \end{aligned}$$

- Choose $\{\hat{a}_k\}$ to minimize the following empirical counterpart (as in Birgé and Massart 97):

$$-\frac{2}{N} \sum_{i=1}^N \sum_k a_k \varphi_k(Y_i, X_i) + \left\| \sum_k a_k \varphi_k(Y, X) \right\|_2^2.$$

- Quadratic problem (least squares). However: possibly high-dimensional \Rightarrow need to penalize.

SPADES

- Bunea, Tsybakov, Wegkamp and Barbu (AS 10) use a Lasso penalty, and minimize:

$$-\frac{2}{N} \sum_{i=1}^N \sum_k a_k \varphi_k(Y_i, X_i) + \left\| \sum_k a_k \varphi_k(Y, X) \right\|_2^2 + \lambda_N \sum_k |a_k|.$$

- Quadratic programming problem. There exist fast routines which compute the Lasso “path”, for all values of λ_N (see the webpage of R. Tibshirani).
- Bunea *et al.* apply this idea to density estimation and discrete mixtures, and study some properties of their SPArse Density ESTimator.
- This discussion suggests that their idea can be used more generally in latent variables models.

Numerical example

$$\begin{cases} Y &= \frac{1}{\theta} \log (1 + \exp (\theta X^*)) + V \\ X &= X^* + U, \end{cases}$$

where V is standard normal, and U is independent standard normal (known).

- DGP: $X^* \sim \mathcal{N}(0, 1)$, $\theta = 1/2$. Use ($\phi =$ standard normal pdf):

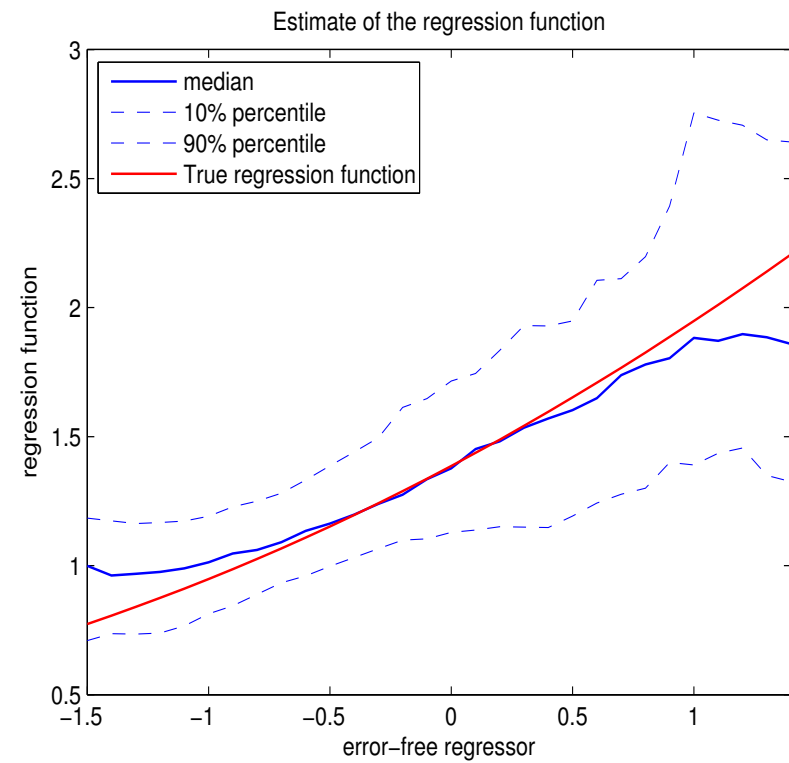
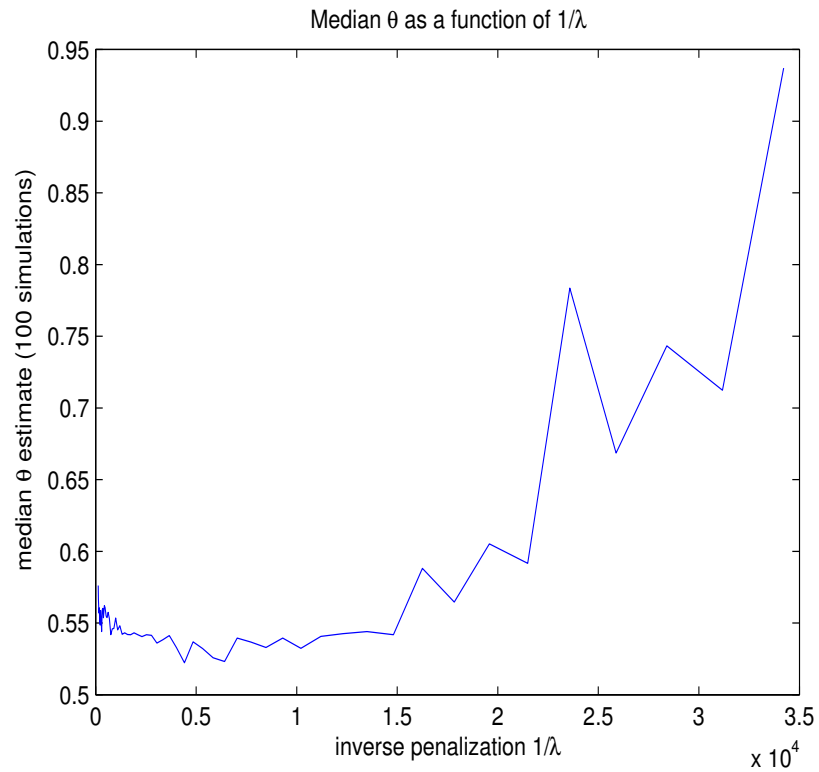
$$\Psi_{k,\ell}(Y, X^*) = \frac{1}{\sigma_k} \phi\left(\frac{Y - \mu_k}{\sigma_k}\right) \times \frac{1}{\sigma_\ell} \phi\left(\frac{X^* - \mu_\ell}{\sigma_\ell}\right), \quad (k, \ell) \in \{1, \dots, 25\}^2,$$

where $\mu_k \sim \mathcal{N}(0, 1)$, and $\sigma_k \sim \frac{1}{10} + \chi_1^2$.

- May estimate θ *ex-post*, as:

$$\hat{\theta} = \operatorname{argmin}_{\theta} \sum_{k,\ell} \hat{a}_{k,\ell} \iint \left(Y - \frac{1}{\theta} \log (1 + \exp (\theta X^*)) \right)^2 \Psi_{k,\ell}(Y, X^*) dY dX^*.$$

Measurement error model ($N = 5000$, true $\theta = 1/2$)



Model:

$$\begin{cases} Y &= \frac{1}{\theta} \log(1 + \exp(\theta X^*)) + V \\ X &= X^* + U, \end{cases}$$

Open issue 1: multiple unknown distributions

- Example: two independent repeated measures

$$\begin{cases} X &= X^* + U \\ \tilde{X} &= X^* + W. \end{cases}, \quad U, W, X^* \text{ mutually independent.}$$

$$\begin{aligned} f(X, \tilde{X}) &= \int f_U(X - X^*) f_W(\tilde{X} - X^*) f_{X^*}(X^*) dX^* \\ &= \underbrace{\sum_{k,\ell,m} a_k b_\ell c_m \int \zeta_\ell(X - X^*) \nu_m(\tilde{X} - X^*) \psi_k(X^*) dX^*}_{\equiv \varphi_{k,\ell,m}(X, \tilde{X})}. \end{aligned}$$

- SPADES: minimize

$$\begin{aligned} -\frac{2}{N} \sum_{i=1}^N \sum_{k,\ell,m} a_k b_\ell c_m \varphi_{k,\ell,m}(X_i, \tilde{X}_i) &+ \left\| \sum_{k,\ell,m} a_k b_\ell c_m \varphi_{k,\ell,m}(X, \tilde{X}) \right\|_2^2 \\ &+ \lambda_N (\sum_k |a_k| + \sum_\ell |b_\ell| + \sum_m |c_m|). \end{aligned}$$

\Rightarrow polynomial programming problem.

Open issue 2: common parameters

- Suppose that $f(X|X^*; \theta)$ is known up to a low dimensional θ .

Ex: structural parameters in economic models with unobserved individual heterogeneity.

- The penalized minimization of the ℓ^2 distance suggests computing:

$$\operatorname{argmin}_{\theta, \{a_k\}} -\frac{2}{N} \sum_{i=1}^N \sum_k a_k \varphi_k(Y_i, X_i; \theta) + \left\| \sum_k a_k \varphi_k(Y, X; \theta) \right\|_2^2 + \lambda_N \sum_k |a_k|.$$

- Computational challenge: nonlinear programming. A possibility is to iterate between Newton optimization (θ) and Lasso ($\{a_k\}$).

- Statistical challenge: at which rate should one let λ_N tend to zero for $\hat{\theta}$ to be root- N consistent and asymptotically normal?

Open issue 3: conditioning covariates

- Presence of covariates Z in $f(X^*|Z) \Rightarrow$ possible curse of dimensionality. Ex: panel data applications.

- In the measurement error example: model the joint density:

$$f(Y, X^*, Z) = \sum_k a_k \psi_k(Y, X^*, Z).$$

- Proposed solution: apply SPADES to the problem of estimating the joint density $f(Y, X, Z)$.

- Difficulty: very high dimensionality of the dictionary $\{\psi_k(Y, X^*, Z)\}$.

\Rightarrow Statistical and computational issues (very large p in Chernozhukov's notation).

Concluding remarks

- Latent variables models are high-dimensional. Penalization methods provide interesting candidate estimators in those models.
- ℓ^2 -distance minimization with ℓ^1 (Lasso) penalty, recently proposed in the context of density estimation, can be extended to more general LVM.
- The true distributions need not be sparse. Rather, the requirement is that some sparse approximations fit those distributions well.
- We have only listed a few of the challenges for application to LVM.
⇒ Lots of interesting research to come!