On conditional moments of high-dimensional random vectors given lower-dimensional projections

Hannes Leeb (University of Vienna)
(with Ivana Milovic and Lukas Steinberger)

Statistical Inference for Large Scale Data
Simon Fraser University
April 20, 2015

universität wien

Sparse models

Among a large collection of potentially important factors or explanatory variables, a sparse (working) model uses only a small subset. (Economics, finance, genomics, proteomics, social sciences, etc.)

Why do you fit sparse models?

- Because I assume, or know, that the true model is sparse.
- Because I do not have enough data to fit the full model.

SPARSE MODELS

Among a large collection of potentially important factors or explanatory variables, a sparse (working) model uses only a small subset. (Economics, finance, genomics, proteomics, social sciences, etc.)

Why do you fit sparse models?

- Because I assume, or know, that the true model is sparse.
- Because I do not have enough data to fit the full model.

SPARSE MODELS

Among a large collection of potentially important factors or explanatory variables, a sparse (working) model uses only a small subset. (Economics, finance, genomics, proteomics, social sciences, etc.)

Why do you fit sparse models?

- Because I assume, or know, that the true model is sparse.
- Because I do not have enough data to fit the full model.

SPARSE MODELS

Among a large collection of potentially important factors or explanatory variables, a sparse (working) model uses only a small subset. (Economics, finance, genomics, proteomics, social sciences, etc.)

Why do you fit sparse models?

- Because I assume, or know, that the true model is sparse.
- Because I do not have enough data to fit the full model.

SPARSE MODELS

Among a large collection of potentially important factors or explanatory variables, a sparse (working) model uses only a small subset. (Economics, finance, genomics, proteomics, social sciences, etc.)

Why do you fit sparse models?

- Because I assume, or know, that the true model is sparse.
- Because I do not have enough data to fit the full model.

SPARSE MODELS: WITH AND WITHOUT SPARSITY

For illustration, consider, as the full model, the homoskedastic linear model

$$Y \quad = \quad X\theta + U,$$

were X is $n \times d$ with $n \ll d$.

Moreover, assume that $\mathbb{E}[U] = 0$, that $\mathbb{E}[UU'] = \sigma^2 I_n$, and that U independent of X.

Partition $X = (X_1 : X_2)$ where X_1 is $n \times p$ with $p < n$, partition $\theta' = (\theta_1', \theta_2')$ conformably, and consider the (sparse) submodel where Y is regressed on X_1.

The submodel is *correct* if $\theta_2 = 0$. Then we have $\mathbb{E}[X\theta||X_1] = X_1\theta_1$ and $\mathbb{V}ar[X\theta||X_1] = 0$.

The submodel *'correct'* if $\mathbb{E}[X\theta||X_1]$ is linear in X_1 and $\mathbb{V}ar[X\theta||X_1]$ is a multiple of the identity.

SPARSE MODELS: WITH AND WITHOUT SPARSITY

For illustration, consider, as the full model, the homoskedastic linear model

$$Y \quad = \quad X\theta + U,$$

were X is $n \times d$ with $n \ll d$.

Moreover, assume that $\mathbb{E}[U] = 0$, that $\mathbb{E}[UU'] = \sigma^2 I_n$, and that U independent of X.

Partition $X = (X_1 : X_2)$ where X_1 is $n \times p$ with $p < n$, partition $\theta' = (\theta_1', \theta_2')$ conformably, and consider the (sparse) submodel where Y is regressed on X_1.

The submodel is *correct* if $\theta_2 = 0$. Then we have $\mathbb{E}[X\theta||X_1] = X_1\theta_1$ and $\mathbb{V}ar[X\theta||X_1] = 0$.

The submodel *'correct'* if $\mathbb{E}[X\theta||X_1]$ is linear in X_1 and $\mathbb{V}ar[X\theta||X_1]$ is a multiple of the identity.

SPARSE MODELS: WITH AND WITHOUT SPARSITY

For illustration, consider, as the full model, the homoskedastic linear model

$$Y \quad = \quad X\theta + U,$$

were X is $n \times d$ with $n \ll d$.

Moreover, assume that $\mathbb{E}[U] = 0$, that $\mathbb{E}[UU'] = \sigma^2 I_n$, and that U independent of X.

Partition $X = (X_1 : X_2)$ where X_1 is $n \times p$ with $p < n$, partition $\theta' = (\theta_1', \theta_2')$ conformably, and consider the (sparse) submodel where Y is regressed on X_1.

The submodel is *correct* if $\theta_2 = 0$. Then we have $\mathbb{E}[X\theta \| X_1] = X_1\theta_1$ and $\mathbb{V}ar[X\theta \| X_1] = 0$.

The submodel *'correct'* if $\mathbb{E}[X\theta \| X_1]$ is linear in X_1 and $\mathbb{V}ar[X\theta \| X_1]$ is a multiple of the identity.

SPARSE MODELS: WITH AND WITHOUT SPARSITY

For illustration, consider, as the full model, the homoskedastic linear model

$$Y \quad = \quad X\theta + U,$$

were X is $n \times d$ with $n \ll d$.

Moreover, assume that $\mathbb{E}[U] = 0$, that $\mathbb{E}[UU'] = \sigma^2 I_n$, and that U independent of X.

Partition $X = (X_1 : X_2)$ where X_1 is $n \times p$ with $p < n$, partition $\theta' = (\theta_1', \theta_2')$ conformably, and consider the (sparse) submodel where Y is regressed on X_1.

The submodel is *correct* if $\theta_2 = 0$. Then we have $\mathbb{E}[X\theta \| X_1] = X_1\theta_1$ and $\mathbb{V}ar[X\theta \| X_1] = 0$.

The submodel *'correct'* if $\mathbb{E}[X\theta \| X_1]$ is linear in X_1 and $\mathbb{V}ar[X\theta \| X_1]$ is a multiple of the identity.

Sparse models: With and without sparsity

Recall that the data-generating model is

$$Y = X\theta + U.$$

If the working model is *correct*, then

$$Y = X_1\theta_1 + U.$$

If the working model is *'correct'*, then

$$Y = X_1\beta + V,$$

for some $\beta \in \mathbb{R}^p$, and with $\mathbb{E}[V\|X_1] = 0$ and $\mathbb{V}ar[V\|X_1] = \varsigma^2 I_n$.
If the working model 'correct' but not correct, we have $\varsigma^2 > \sigma^2$, because

$$V = U + (X\theta - \mathbb{E}[X\theta\|X_1])$$

and hence

$$\varsigma^2 = \sigma^2 + \mathbb{V}ar[X\theta\|X_1].$$

SPARSE MODELS: WITH AND WITHOUT SPARSITY

Recall that the data-generating model is

$$Y = X\theta + U.$$

If the working model is *correct*, then

$$Y = X_1\theta_1 + U.$$

If the working model is *'correct'*, then

$$Y = X_1\beta + V,$$

for some $\beta \in \mathbb{R}^p$, and with $\mathbb{E}[V\|X_1] = 0$ and $\mathbb{V}ar[V\|X_1] = \varsigma^2 I_n$.
If the working model 'correct' but not correct, we have $\varsigma^2 > \sigma^2$, because

$$V = U + (X\theta - \mathbb{E}[X\theta\|X_1])$$

and hence

$$\varsigma^2 = \sigma^2 + \mathbb{V}ar[X\theta\|X_1].$$

Sparse models: With and without sparsity

Recall that the data-generating model is

$$Y \quad = \quad X\theta + U.$$

If the working model is *correct*, then

$$Y \quad = \quad X_1\theta_1 + U.$$

If the working model is *'correct'*, then

$$Y \quad = \quad X_1\beta + V,$$

for some $\beta \in \mathbb{R}^p$, and with $\mathbb{E}[V\|X_1] = 0$ and $\mathbb{V}ar[V\|X_1] = \varsigma^2 I_n$.

If the working model 'correct' but not correct, we have $\varsigma^2 > \sigma^2$, because

$$V \quad = \quad U + (X\theta - \mathbb{E}[X\theta\|X_1])$$

and hence

$$\varsigma^2 \quad = \quad \sigma^2 + \mathbb{V}ar[X\theta\|X_1].$$

Sparse models: With and without sparsity

Recall that the data-generating model is

$$Y \quad = \quad X\theta + U.$$

If the working model is *correct*, then

$$Y \quad = \quad X_1\theta_1 + U.$$

If the working model is *'correct'*, then

$$Y \quad = \quad X_1\beta + V,$$

for some $\beta \in \mathbb{R}^p$, and with $\mathbb{E}[V\|X_1] = 0$ and $\mathbb{V}ar[V\|X_1] = \varsigma^2 I_n$.
If the working model 'correct' but not correct, we have $\varsigma^2 > \sigma^2$, because

$$V \quad = \quad U + (X\theta - \mathbb{E}[X\theta\|X_1])$$

and hence

$$\varsigma^2 \quad = \quad \sigma^2 + \mathbb{V}ar[X\theta\|X_1].$$

'CORRECT' SUBMODELS

Consider one observation from the full model, i.e.,

$$y \quad = \quad x'\theta + u,$$

and consider the submodel using only x_1 (where $x' = (x_1', x_2')$).

The submodel is 'correct' if θ, x, and x_1 satisfy

- $\mathbb{E}[\theta' x || x_1]$ is linear in x_1 and
- $\mathbb{V}ar[\theta' x || x_1]$ is constant in x_1.

'CORRECT' SUBMODELS

Consider one observation from the full model, i.e.,

$$y \quad = \quad x'\theta + u,$$

and consider the submodel using only x_1 (where $x' = (x_1', x_2')$).

The submodel is 'correct' if θ, x, and x_1 satisfy

- $\mathbb{E}[\theta' x \| x_1]$ is linear in x_1 and
- $\mathbb{V}ar[\theta' x \| x_1]$ is constant in x_1.

The submodel is 'correct' *irrespective of* θ if x and x_1 satisfy

- $\mathbb{E}[x \| x_1]$ is linear in x_1 and
- $\mathbb{V}ar[x \| x_1]$ is constant in x_1.

'CORRECT' SUBMODELS

Consider one observation from the full model, i.e.,

$$y = x'\theta + u,$$

and consider the submodel using only x_1 (where $x' = (x_1', x_2')$).

The submodel is 'correct' if θ, x, and x_1 satisfy

- $\mathbb{E}[\theta' x \| x_1]$ is linear in x_1 and
- $\mathbb{V}ar[\theta' x \| x_1]$ is constant in x_1.

The submodel is 'correct' *irrespective of* θ if x and x_1 satisfy

- $\mathbb{E}[x \| x_1]$ is linear in x_1 and
- $\mathbb{V}ar[x \| x_1]$ is constant in x_1.

Under the latter condition, standard methods can be used to perform inference based on the submodel.

'CORRECT' SUBMODELS

Consider one observation from the full model, i.e.,

$$y \quad = \quad x'\theta + u,$$

and consider the submodel using only x_1 (where $x' = (x_1', x_2')$).

The submodel is 'correct' if θ, x, and x_1 satisfy

- $\mathbb{E}[\theta' x \| x_1]$ is linear in x_1 and
- $\mathbb{V}ar[\theta' x \| x_1]$ is constant in x_1.

The submodel is 'correct' *irrespective of* θ if x and x_1 satisfy

- $\mathbb{E}[x \| x_1]$ is linear in x_1 and
- $\mathbb{V}ar[x \| x_1]$ is constant in x_1.

The latter condition is restrictive: The Gaussian distribution is, in a sense, characterized by linear conditional means and constant conditional variances (Eaton 1986, Bryc 1995).

Informal summary: Most submodels are 'correct'

Recall that a submodel with regressors x_1 is 'correct' if $\mathbb{E}[x||x_1]$ is linear and $\mathbb{V}ar[x||x_1]$ is constant in x_1.

We show:

For a large class of non-Gaussian distributions, *most* conditional means are *approximately linear* and *most* conditional variances are *approximately constant*.

Our approximation errors go to zero provided that the dimension of the conditioning vector (submodel) is small relative to the dimension of the overall vector (full model). (Large-dimension asymptotics)

Informal summary: Most submodels are 'correct'

Recall that a submodel with regressors x_1 is 'correct' if $\mathbb{E}[x||x_1]$ is linear and $\mathbb{V}ar[x||x_1]$ is constant in x_1.

We show:

For a large class of non-Gaussian distributions, *most* conditional means are *approximately linear* and *most* conditional variances are *approximately constant*.

Our approximation errors go to zero provided that the dimension of the conditioning vector (submodel) is small relative to the dimension of the overall vector (full model). (Large-dimension asymptotics)

Informal summary: Most submodels are 'correct'

Recall that a submodel with regressors x_1 is 'correct' if $\mathbb{E}[x||x_1]$ is linear and $\mathbb{V}ar[x||x_1]$ is constant in x_1.

We show:

For a large class of non-Gaussian distributions, *most* conditional means are *approximately linear* and *most* conditional variances are *approximately constant*.

Our approximation errors go to zero provided that the dimension of the conditioning vector (submodel) is small relative to the dimension of the overall vector (full model). (Large-dimension asymptotics)

Informal summary: Most submodels are 'correct'

Recall that a submodel with regressors x_1 is 'correct' if $\mathbb{E}[x||x_1]$ is linear and $\mathbb{V}ar[x||x_1]$ is constant in x_1.

We show:

For a large class of non-Gaussian distributions, *most* conditional means are *approximately linear* and *most* conditional variances are *approximately constant*.

Our approximation errors go to zero provided that the dimension of the conditioning vector (submodel) is small relative to the dimension of the overall vector (full model). (Large-dimension asymptotics)

Note: On a technical level, our results should be compared to those of Hall and Li (1993) and also to those of Diaconis and Freedman (1984).

Informal summary: Most submodels are 'correct'

Recall that a submodel with regressors x_1 is 'correct' if $\mathbb{E}[x||x_1]$ is linear and $\mathbb{V}ar[x||x_1]$ is constant in x_1.

We show:

For a large class of non-Gaussian distributions, *most* conditional means are *approximately linear* and *most* conditional variances are *approximately constant*.

Our approximation errors go to zero provided that the dimension of the conditioning vector (submodel) is small relative to the dimension of the overall vector (full model). (Large-dimension asymptotics)

Note: Conceptually, our approach is similar to that of Berk, Brown, Buja, Zhang and Zhao (2013), Genovese and Wasserman (2008), or Leeb (2009).

Quantities of interest

For each dimension d, consider a random d-vector Z that has a Lebesgue density and that is standardized so that $\mathbb{E}Z = 0$ and $\mathbb{E}ZZ' = I_d$. We also impose a technical condition (c) that will be introduced and discussed later.

Our results are asymptotic as $d \to \infty$. We allow for $p \to \infty$ subject to $p \ll d$.

Consider a $d \times p$ matrix B with orthonormal columns. Conditional on $B'Z$, the mean of Z is linear and the variance of Z is constant in the conditioning variables, if both

$$\left\| \mathbb{E}[Z \| B'Z = x] - Bx \right\| = 0$$

$$\left\| \mathbb{E}[ZZ' \| B'Z = x] - (I_d + B(xx' - I_p)B') \right\| = 0$$

hold for each $x \in \mathbb{R}^p$. The standardizations of Z and B are inconsequential.

Quantities of interest

For each dimension d, consider a random d-vector Z that has a Lebesgue density and that is standardized so that $\mathbb{E}Z = 0$ and $\mathbb{E}ZZ' = I_d$. We also impose a technical condition (c) that will be introduced and discussed later.

Our results are asymptotic as $d \to \infty$. We allow for $p \to \infty$ subject to $p \ll d$.

Consider a $d \times p$ matrix B with orthonormal columns. Conditional on $B'Z$, the mean of Z is linear and the variance of Z is constant in the conditioning variables, if both

$$\begin{aligned} \left\| \mathbb{E}[Z \| B'Z = x] - Bx \right\| &= 0 \\ \left\| \mathbb{E}[ZZ' \| B'Z = x] - (I_d + B(xx' - I_p)B') \right\| &= 0 \end{aligned}$$

hold for each $x \in \mathbb{R}^p$. The standardizations of Z and B are inconsequential.

Quantities of interest

For each dimension d, consider a random d-vector Z that has a Lebesgue density and that is standardized so that $\mathbb{E}Z = 0$ and $\mathbb{E}ZZ' = I_d$. We also impose a technical condition (c) that will be introduced and discussed later.

Our results are asymptotic as $d \to \infty$. We allow for $p \to \infty$ subject to $p \ll d$.

Consider a $d \times p$ matrix B with orthonormal columns. Conditional on $B'Z$, the mean of Z is linear and the variance of Z is constant in the conditioning variables, if both

$$\begin{aligned} \left\| \mathbb{E}[Z \| B'Z = x] - Bx \right\| &= 0 \\ \left\| \mathbb{E}[ZZ' \| B'Z = x] - (I_d + B(xx' - I_p)B') \right\| &= 0 \end{aligned}$$

hold for each $x \in \mathbb{R}^p$. The standardizations of Z and B are inconsequential.

Two preliminary results

Consider the conditions

$$\begin{aligned} \left\| \mathbb{E}[Z \| B'Z = x] - Bx \right\| &= 0 \\ \left\| \mathbb{E}[ZZ' \| B'Z = x] - (I_d + B(xx' - I_p)B') \right\| &= 0 \end{aligned}$$

for each $x \in \mathbb{R}^p$.

Two preliminary results

Consider the conditions

$$\left\| \mathbb{E}[Z \| B'Z = x] - Bx \right\| = 0$$

$$\left\| \mathbb{E}[ZZ' \| B'Z = x] - (I_d + B(xx' - I_p)B') \right\| = 0$$

for each $x \in \mathbb{R}^p$.

Proposition 1 (Hall & Li 1993):

Set $p = 1$ and impose condition (c). For each fixed $x \in \mathbb{R}$ and each $t > 0$, we have

$$\nu_{d,1} \left\{ B \in \mathbb{R}^d : \left\| \mathbb{E}[Z \| B'Z = x] - Bx \right\| > t \right\} \overset{d \to \infty}{\longrightarrow} 0,$$

where $\nu_{d,1}$ denotes the uniform distribution on the unit sphere in $\mathbb{R}^d$.

Two preliminary results

Consider the conditions

$$\begin{aligned} \left\| \mathbb{E}[Z \| B'Z = x] - Bx \right\| &= 0 \\ \left\| \mathbb{E}[ZZ' \| B'Z = x] - (I_d + B(xx' - I_p)B') \right\| &= 0 \end{aligned}$$

for each $x \in \mathbb{R}^p$.

Proposition 2 (L, 2013):

Set $p = 1$ and impose condition (c). For each fixed $x \in \mathbb{R}$ and each $t > 0$, we have

$$\nu_{d,1} \left\{ B \in \mathbb{R}^d : \left\| \mathbb{E}[ZZ' \| B'Z = x] - (I_d + B(x^2 - 1)B') \right\| > t \right\} \overset{d \to \infty}{\longrightarrow} 0.$$

Two preliminary results

Consider the conditions

$$\left\| \mathbb{E}[Z \| B'Z = x] - Bx \right\| = 0$$

$$\left\| \mathbb{E}[ZZ' \| B'Z = x] - (I_d + B(xx' - I_p)B') \right\| = 0$$

for each $x \in \mathbb{R}^p$.

If the last two propositions apply, then the left-hand sides in the preceding display are small, for *most* B*'s* but only for *fixed* $x \in \mathbb{R}$, provided only that d is large. To show that this also applies for *most* x*'s*, we now replace x by $B'Z$, and we now also allow for $p > 1$.

Main result, qualitative version

Write $\mathcal{V}_{d,p}$ for the collection of all $d \times p$ matrices B with orthonormal columns (Stiefel manifold), and denote the corresponding Haar measure by $\nu_{d,p}$.

Main result, qualitative version

Write $\mathcal{V}_{d,p}$ for the collection of all $d \times p$ matrices B with orthonormal columns (Stiefel manifold), and denote the corresponding Haar measure by $\nu_{d,p}$.

Theorem 1 (L, 2013; Steinberger & L, 2014):

For each fixed d, consider a random d-vector Z that has a Lebesgue density and that is standardized to that $\mathbb{E}Z = 0$ and $\mathbb{E}ZZ' = I_d$. Suppose that condition (c) is satisfied, and suppose that $p = o(\log d)$. Then there are Borel sets $G_{d,p} \subseteq \mathcal{V}_{d,p}$ so that $\nu_{d,p}(G_{d,p}) \to 1$ as $d \to \infty$, and so that

$$\sup_{B \in G_{d,p}} \mathbb{P}\left(\left\| \mathbb{E}[Z \| B'Z] - BB'Z \right\| > t \right) \qquad \text{and} \tag{1}$$

$$\sup_{B \in G_{d,p}} \mathbb{P}\left(\left\| \mathbb{E}[ZZ' \| B'Z] - (I_d + B(B'ZZ'B - I_p)B') \right\| > t \right) \tag{2}$$

converge to zero as $d \to \infty$ for each $t > 0$.

Main result, quantitative version

Theorem 2 (Steinberger & L, 2014):

For fixed d, consider a random d-vector Z that has a Lebesgue density and that is standardized such that $\mathbb{E}Z = 0$ and $\mathbb{E}ZZ' = I_d$. Suppose that condition (c) is satisfied. Then, for each $p < d$, there is a Borel set $G_{d,p} \subseteq \mathcal{V}_{d,p}$ so that both (1) and (2) are bounded by

$$\frac{1}{t} d^{-\frac{\xi_\epsilon}{2}} + \frac{p}{\log d} \frac{2\gamma}{5\xi_\epsilon}$$

for each $t > 0$, and such that

$$\nu_{d,p}(G_{d,p}^c) \quad \leq \quad 2\kappa\, d^{-\frac{\xi_\epsilon}{2}\left(1 - \frac{p}{\log d} \frac{2\gamma}{\xi_\epsilon}\right)}.$$

Here, ξ_ϵ is given by $\xi_\epsilon = \min\{\xi, \epsilon/2 + 1/4, 1/2\}/5$, and ϵ, ξ, γ and κ are constants derived from condition (c) that do not depend on p or d.

Condition (c) imposed in Theorems 1 & 2:

Condition (c) – the simple version:

The components of Z are independent with bounded marginal moments and bounded marginal densities of sufficiently high order (where the bounds do not depend on d).

Condition (c) imposed in Theorems 1 & 2:

A more general version of condition (c) relies on a set of finite-sample moment bounds involving the Gram-matrix

$$S_k \quad = \quad (Z_i' Z_j / d)_{i,j=1}^{k}$$

for Z_i, $i = 1, \ldots, k$, being i.i.d. copies of Z.

Condition (c) imposed in Theorems 1 & 2:

- **(b1)** For fixed $k \in \mathbb{N}$, there are constants $\epsilon \in [0, 1/2]$, $\alpha \geq 1$, $\beta > 0$, and $\xi \in (0, 1/2]$, such that the following holds true:
  - **(a)** $\mathbb{E}\|\sqrt{d}(S_k - I_k)\|^{2k+1+\epsilon} \leq \alpha$.
  - **(b)** For any monomial $G = G(S_k - I_k)$ in the elements of $S_k - I_k$, whose degree g satisfies $g \leq 2k$, we have $|d^{g/2}\mathbb{E}G - 1| \leq \beta/d^{\xi}$ if G consists only of quadratic terms in elements above the diagonal, and $|d^{g/2}\mathbb{E}G| \leq \beta/d^{\xi}$ if G contains a linear term.
  - **(c)** Consider two monomials $G = G(S_k - I_k)$ and $H = H(S_k - I_k)$ of degree g and h, respectively, in the elements of $S_k - I_k$. If G is given by $Z_1'Z_2Z_2'Z_3 \ldots Z_{g-1}'Z_gZ_g'Z_1/d^g$, if H depends at least on those Z_i's with $i \leq g$, and if $2 \leq h < g \leq k$, then $|d^g\mathbb{E}GH| \leq \beta/d^{\xi}$.
- **(b2)** For fixed $k \in \mathbb{N}$, there is a constant $D \geq 1$, such that the following holds true: If R is an orthogonal $d \times d$ matrix, then a marginal density of the first $d - k + 1$ components of RZ is bounded by $\binom{d}{k-1}^{1/2} D^{d-k+1}$.

Condition (c) imposed in Theorems 1 & 2:

Condition (c) – the general version:

For each d and p under consideration, suppose that (b1) and (b2) hold with $k = 4$, such that the constants ϵ, ξ, α, β, and D in these bounds do not depend on d or p.

The general version of condition (c) allows for dependent components and unbounded marginal moments.

Condition (c) imposed in Theorems 1 & 2:

Condition (c) – the general version:

For each d and p under consideration, suppose that (b1) and (b2) hold with $k = 4$, such that the constants ϵ, ξ, α, β, and D in these bounds do not depend on d or p.

The general version of condition (c) allows for dependent components and unbounded marginal moments.

Condition (c) imposed in Theorems 1 & 2:

Remarks:

- The bound in (b2) appears to be qualitatively different from (b1) in that it does not directly impose restrictions to moments of the standardized Gram matrix $S_k - I_k$. However, (b2) is used only to bound the p-th moment of $\det S_l^{-4(k+1)}$ for $l = 1, \ldots, k$.
- (b1).(a) and (b2) guarantee that the mean of certain functions of Z, and of i.i.d. copies of Z, is bounded. And (b1).(b-c) require that certain moments of Z are close to what they would be in the Gaussian case.
- From a statistical perspective, we note that the moments discussed here can be estimated from a sample of independent copies of Z. Indeed, population means like $\mathbb{E}\|S_k - I_k\|^{2k+1+\epsilon}$, $\mathbb{E}G$, $\mathbb{E}GH$, or $\mathbb{E}\det S_l^{-4p(k+1)}$ as above are readily estimated by appropriate sample means. In this sense, we rely on bounds on quantities that can be estimated from data.

SOME INTUITION

For each d, let Z be a d-vector with i.i.d. components (from the same distribution), each with mean zero and variance 1.

By the central limit theorem, we have

$$\alpha' Z \quad \xrightarrow{w} \quad N(0,1)$$

for unit-vectors α of the form $\alpha' = (1/\sqrt{d}, \ldots, 1/\sqrt{d})$ and as $d \to \infty$.

Diaconis and Freedman (1984) show that the above holds for 'most' unit-vectors α:

$$\nu_{d,1}\left\{\alpha : \; d(\mathcal{L}(\alpha' Z), N(0,1)) > t\right\} \quad \longrightarrow \quad 1$$

as $d \to \infty$ for some metric $d(\cdot, \cdot)$ that 'implies weak convergence'.

SOME INTUITION

For each d, let Z be a d-vector with i.i.d. components (from the same distribution), each with mean zero and variance 1.

By the central limit theorem, we have

$$\alpha' Z \quad \xrightarrow{w} \quad N(0,1)$$

for unit-vectors α of the form $\alpha' = (1/\sqrt{d}, \ldots, 1/\sqrt{d})$ and as $d \to \infty$.

Diaconis and Freedman (1984) show that the above holds for 'most' unit-vectors α:

$$\nu_{d,1}\left\{\alpha : \; d(\mathcal{L}(\alpha' Z), N(0,1)) > t\right\} \quad \longrightarrow \quad 1$$

as $d \to \infty$ for some metric $d(\cdot,\cdot)$ that 'implies weak convergence'.

SOME INTUITION

For each d, let Z be a d-vector with i.i.d. components (from the same distribution), each with mean zero and variance 1.

By the central limit theorem, we have

$$\alpha' Z \quad \xrightarrow{w} \quad N(0,1)$$

for unit-vectors α of the form $\alpha' = (1/\sqrt{d}, \ldots, 1/\sqrt{d})$ and as $d \to \infty$.

Diaconis and Freedman (1984) show that the above holds for 'most' unit-vectors α:

$$\nu_{d,1}\left\{\alpha : \; d(\mathcal{L}(\alpha' Z), N(0,1)) > t\right\} \quad \longrightarrow \quad 1$$

as $d \to \infty$ for some metric $d(\cdot, \cdot)$ that 'implies weak convergence'.

Some intuition

The result of Diaconis and Freedman (1984) also holds for most pairs of unit-vectors α and β. In other words,

$$\mathcal{L}(\alpha' Z, \beta' Z) \quad \approx \quad N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \alpha'\beta \\ \alpha'\beta & 1 \end{pmatrix}\right)$$

for most pairs (α, β) and as $d \to \infty$.

Some intuition

The result of Diaconis and Freedman (1984) also holds for most pairs of unit-vectors α and β. In other words,

$$\mathcal{L}(\alpha' Z, \beta' Z) \quad \approx \quad N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \alpha'\beta \\ \alpha'\beta & 1 \end{pmatrix}\right)$$

for most pairs (α, β) and as $d \to \infty$.

This makes our result plausible. But convergence of distributions typically does not entail convergence of moments or convergence of conditional moments.

SOME INTUITION

The result of Diaconis and Freedman (1984) also holds for most pairs of unit-vectors α and β. In other words,

$$\mathcal{L}(\alpha'Z, \beta'Z) \quad \approx \quad N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \alpha'\beta\\ \alpha'\beta & 1\end{pmatrix}\right)$$

for most pairs (α, β) and as $d \to \infty$.

Idea 1:
For most pairs of unit-vectors (α, β) and as $d \to \infty$, we have

$$\begin{aligned}\mathbb{E}[\alpha'Z \| \beta'Z] &\approx \alpha'\beta\beta'Z,\\ \mathbb{V}ar[\alpha'Z \| \beta'Z] &\approx 1 - (\alpha'\beta)^2\end{aligned}$$

SOME INTUITION

The result of Diaconis and Freedman (1984) also holds for most pairs of unit-vectors α and β. In other words,

$$\mathcal{L}(\alpha'Z, \beta'Z) \quad \approx \quad N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \alpha'\beta\\ \alpha'\beta & 1\end{pmatrix}\right)$$

for most pairs (α, β) and as $d \to \infty$.

Idea 1:
For most pairs of unit-vectors (α, β) and as $d \to \infty$, we have

$$\begin{aligned}\mathbb{E}[\alpha'Z \| \beta'Z] &\approx \alpha'\beta\beta'Z,\\ \mathbb{V}ar[\alpha'Z \| \beta'Z] &\approx 1 - (\alpha'\beta)^2\end{aligned}$$

But $\alpha'\beta \approx 0$ for most pairs (α, β) and as $d \to \infty$.

Some intuition

The result of Diaconis and Freedman (1984) also holds for most pairs of unit-vectors α and β. In other words,

$$\mathcal{L}(\alpha' Z, \beta' Z) \quad \approx \quad N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \alpha'\beta \\ \alpha'\beta & 1 \end{pmatrix}\right)$$

for most pairs (α, β) and as $d \to \infty$.

Idea 2:
For most unit-vectors β and as $d \to \infty$, we have

$$\begin{aligned} \mathbb{E}[Z \| \beta' Z] &\approx \beta\beta' Z, \\ \mathbb{V}ar[Z \| \beta' Z] &\approx I_d - \beta\beta'. \end{aligned}$$

Some intuition

The result of Diaconis and Freedman (1984) also holds for most pairs of unit-vectors α and β. In other words,

$$\mathcal{L}(\alpha' Z, \beta' Z) \quad \approx \quad N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \alpha'\beta \\ \alpha'\beta & 1 \end{pmatrix}\right)$$

for most pairs (α, β) and as $d \to \infty$.

Idea 3:
For most matrices B from the Stiefel-manifold $\mathcal{V}_{d,p}$ and as $d \to \infty$, we have

$$\begin{aligned} \mathbb{E}[Z \| B'Z] &\approx BB'Z, \\ \mathbb{V}ar[Z \| \beta' Z] &\approx I_d - BB', \end{aligned}$$

provided that $p = o(\log(d))$. Moreover, we provide explicit error bounds that hold for fixed p and d.

Outlook

- Approximately valid prediction and inference when fitting simple linear submodels to complex data-generating processes; joint with Lukas Steinberger.
- Fast convergence rates for the error bounds in Theorem 2, with applications to model selection and regularization; joint with Ivana Milovic.

References

- Diaconis, P., Freedman, D. (1984): Asymptotics of graphical projection pursuit. *Ann. Statist.* **12** 793–815 (1984)
- Hall, P., Li, K.C.: On almost linearity of low dimensional projections from high dimensional data. *Ann. Statist.* **21**, 867–889 (1993)
- Leeb, H.: On the conditional distribution of low-dimensional projections from high-dimensional data. *Ann. Statist.* **41**, 464–483 (2013)
- Steinberger, L., Leeb, H: On conditional moments of high-dimensional random vectors given lower-dimensional projections, revision for *Bernoulli* in preparation (2014); arXiv:1405.2183.