Matrix and tensor decomposition methods

Peter Hoff

Statistics and Biostatistics, University of Washington
Outline

- Comments
- Extensions
Themes

1. Matrices and arrays of data and parameters
2. Low rank representations
3. Penalization/shrinkage/hierarchical modeling
Matrix and array data and parameters

**Eric:**
- multiple object $\times$ variable datasets, with common objects.

**Jianhua:**
- mortality rate for years $\times$ age
- time of day $\times$ waiting times
- representatives $\times$ votes

**Vadim:**
- $y_{i,j,t}$: person $\times$ time $\times$ voxel
- $\{u_{i,0}(v), u_{i,1}(v)\}$: person $\times$ 2 $\times$ voxel parameter for linear trend in time

Other examples:
- probabilities for contingency tables (Dunson and collaborators)
- interactions for ANOVA decompositions (Volfovsky and Hoff, 2012)
- country $\times$ age $\times$ year $\times$ gender mortality rates (Fosdick and Hoff, 2012)
Matrix and array data and parameters

**Eric:**
- multiple object × variable datasets, with common objects.

**Jianhua:**
- mortality rate for years × age
- time of day × waiting times
- representatives × votes

**Vadim:**
- $y_{i,j,t}$: person × time × voxel
- $\{u_{i,0}(v), u_{i,1}(v)\}$: person × 2 × voxel parameter for linear trend in time

Other examples:
- probabilities for contingency tables (Dunson and collaborators)
- interactions for ANOVA decompositions (Volfovsky and Hoff, 2012)
- country × age × year × gender mortality rates (Fosdick and Hoff, 2012)
Matrix and array data and parameters

Eric:
- multiple object $\times$ variable datasets, with common objects.

Jianhua:
- mortality rate for years $\times$ age
- time of day $\times$ waiting times
- representatives $\times$ votes

Vadim:
- $y_{i,j,t}$: person $\times$ time $\times$ voxel
- $\{u_{i,0}(v), u_{i,1}(v)\}$: person $\times$ 2 $\times$ voxel parameter for linear trend in time

Other examples:
- probabilities for contingency tables (Dunson and collaborators)
- interactions for ANOVA decompositions (Volfovsky and Hoff, 2012)
- country $\times$ age $\times$ year $\times$ gender mortality rates (Fosdick and Hoff, 2012)
Matrix and array data and parameters

Eric:
- multiple object \( \times \) variable datasets, with common objects.

Jianhua:
- mortality rate for years \( \times \) age
- time of day \( \times \) waiting times
- representatives \( \times \) votes

Vadim:
- \( y_{i,j,t} \): person \( \times \) time \( \times \) voxel
- \( \{ u_{i,0}(v), u_{i,1}(v) \} \): person \( \times 2 \ \times \) voxel parameter for linear trend in time

Other examples:
- probabilities for contingency tables (Dunson and collaborators)
- interactions for ANOVA decompositions (Volfovsky and Hoff, 2012)
- country \( \times \) age \( \times \) year \( \times \) gender mortality rates (Fosdick and Hoff, 2012)
Matrix and array data and parameters

Eric:
• multiple object × variable datasets, with common objects.

Jianhua:
• mortality rate for years × age
• time of day × waiting times
• representatives × votes

Vadim:
• $y_{i,j,t}$: person × time × voxel
• $\{u_{i,0}(v), u_{i,1}(v)\}$: person × 2 × voxel parameter for linear trend in time

Other examples:
• probabilities for contingency tables (Dunson and collaborators)
• interactions for ANOVA decompositions (Volfovsky and Hoff, 2012)
• country × age × year × gender mortality rates (Fosdick and Hoff, 2012)
Low-rank representations

Assume mean matrix $\mathbf{M}$ is of the form $\mathbf{UV}^T$ or $\mathbf{USV}^T$

- orthogonal components for $\mathbf{U}$, $\mathbf{V}$ (Eric, Vadim)
- non-orthogonality in the criterion (but maybe orthogonal estimate?)

Questions:

- How to choose the rank?
  - Eric: permutation tests
  - Owen, Perry, Eckles: resampling approaches
  - penalization of the singular values
  - Hoff(2007): prior over ranks
  - Eric Owen, Perry and colleagues: resampling approaches

- How to extend to tensors?
Low-rank representations

Assume mean matrix $M$ is of the form $UV^T$ or $USV^T$

- orthogonal components for $U$, $V$ (Eric, Vadim)
- non-orthogonality in the criterion (but maybe orthogonal estimate?)

Questions:
- How to choose the rank?
  - Eric: permutation tests
  - Owen, Perry, Eckles: resampling approaches
  - penalization of the singular values
  - Hoff(2007): prior over ranks
  - Eric Owen, Perry and colleagues: resampling approaches

- How to extend to tensors?
Penalization and priors

Vadim and Jianhua use L2/quadratic/Gaussian penalties:

\[ ||Y - uv^T||^2 + u^T \Sigma^{-1} u + \cdots \]

Encourages smoothness and similarity according to \( \Sigma \).
Following common practice, we first group all the calls into 6-quarter-hour intervals from 07:00 to midnight. For each interval, we apply the Kaplan-Meier estimator (Kaplan and Meier, 1958) to obtain the survival function of time-willing-to-wait $W$, with which we then calculate the log-odds function of patience $\log\{P(W > w) / P(W \leq w)\}$.

One reason for considering log-odds is that they are interval scale (use the whole real line), which renders them more appropriate for an SVD-based analysis. The final data matrix $X$ consists — for each quarter hour interval — of the evaluation so forth the log-odds function at the seconds 11, 12, ..., 200 for the waiting times. Hence the size of $X$ is $68 \times 190$, where the rows are indexed by 15-minute time-of-day intervals, and the columns are indexed by waiting times in seconds for all seconds from 11 to 200.

The regularized SVD yields the following model of the log-odds as a function of time-of-day $t$ and time-willing-to-wait $w$,

$$X(t, w) = d_1 U_1(t) V_1(w) + \ldots + d_q U_q(t) V_q(w) + \epsilon(t, w),$$

(30)

where $U_i(\cdot)$ and $V_i(\cdot)$ are smooth in time-of-day and time-willing-to-wait, respectively.

Figure 4 compares the first pair of components between plain and regularized SVDs. In Panel (a) the regularized singular curve reveals an interesting double-dip pattern of log-odds as a function of time-of-day. The function decreases from the 26th.
Identifiability issues

\[ ||Y - uv^T||^2 + ||v||^2 u^T \Sigma_u^{-1} u + ||u||^2 v^T \Sigma_v^{-1} v + v^T \Sigma_v^{-1} vu^T \Sigma_u^{-1} u \]

**Interpretation:**

- \( ||v||^2 u^T \Sigma_u^{-1} u \): \( v \) big \( \rightarrow \) \( u \) small.
- \( v^T \Sigma_v^{-1} vu^T \Sigma_u^{-1} u \): not clear (at least before the talk) - the “rounder” \( v \) is the smoother \( u \) must be?

What about imposing scale constraints - \( u^T u = 1 \)?

- \( u^T \Sigma_u^{-1} u \), \( u \) unconstrained corresponds to a normal prior;
- \( u^T \Sigma_u^{-1} u \), \( u \) constrained corresponds to a Bingham distribution prior.

Hoff(2009) - example with a uniform prior/Bingham posterior distribution for \( u \) in a binary probit network model.
Identifiability issues

\[ ||Y - uv^T||^2 + ||v||^2 u^T \Sigma^{-1}_u u + ||u||^2 v^T \Sigma^{-1}_v v + v^T \Sigma^{-1}_v vu^T \Sigma^{-1}_u u \]

**Interpretation:**

- \( ||v||^2 u^T \Sigma^{-1}_u u \): \( v \) big \( \rightarrow \) \( u \) small.
- \( v^T \Sigma^{-1}_v vu^T \Sigma^{-1}_u u \): not clear (at least before the talk) - the “rouglier” \( v \) is the smoother \( u \) must be?

What about imposing scale constraints - \( u^T u = 1 \)?

- \( u^T \Sigma^{-1}_u u \), \( u \) unconstrained corresponds to a normal prior;
- \( u^T \Sigma^{-1}_u u \), \( u \) constrained corresponds to a Bingham distribution prior.

Hoff(2009) - example with a uniform prior/Bingham posterior distribution for \( u \) in a binary probit network model.
Identifiability issues

\[ \|Y - uv^T\|^2 + \|v\|^2 u^T \Sigma_u^{-1} u + \|u\|^2 v^T \Sigma_v^{-1} v + v^T \Sigma_v^{-1} vu^T \Sigma_u^{-1} u \]

**Interpretation:**

- \( \|v\|^2 u^T \Sigma_u^{-1} u : v \text{ big } \rightarrow u \text{ small.} \)
- \( v^T \Sigma_v^{-1} vu^T \Sigma_u^{-1} u : \text{not clear (at least before the talk) - the “rougheer” } v \text{ is the smoother } u \text{ must be?} \)

What about imposing scale constraints - \( u^T u = 1 ? \)

- \( u^T \Sigma_u^{-1} u, \text{ u unconstrained corresponds to a normal prior; } \)
- \( u^T \Sigma_u^{-1} u, \text{ u constrained corresponds to a Bingham distribution prior. } \)

Hoff(2009) - example with a uniform prior/Bingham posterior distribution for \( u \) in a binary probit network model.
Sparsity penalties/priors

\[ \| \mathbf{Y} - \mathbf{u} \mathbf{v}^T \|^2 + \lambda_u \sum |u_i| + \lambda_v \sum |v_j| \]

Less clear on the interpretation here:

- \( u_i = 0 \) wipes out the whole row;
- maybe similar to plaid models?
- Allen TR for extension to arrays.
General comments

- extension to higher-order arrays
- dimension selection and regularization
- extensions to accommodate non-normal data
Generalizing the SVD, part 1

**Mean model:**

\[ Y = M + E \]

**Matrices:**

\[
\text{rank}(M = r) \iff M = \sum_{r=1}^{R} s_r \ u_r v_r^T \\
= \sum_{r=1}^{R} s_r \ u_r \circ v_r = USV^T
\]

**Arrays:**

\[
\text{rank}(M = r) \iff M = \sum_{r=1}^{R} s_r \ u_r \circ v_r \circ w_r
\]

“CP” model (Harshman 1970, Carrol and Chang 1970)
Hoff 2011: Hierarchical Bayes approach
Generalizing the SVD, part 1

Mean model:

\[ Y = M + E \]

Matrices:

\[
\text{rank}(M = r) \Leftrightarrow M = \sum_{r=1}^{R} s_r \, u_r v_r^T \\
= \sum_{r=1}^{R} s_r \, u_r \circ v_r = USV^T
\]

Arrays:

\[
\text{rank}(M = r) \Leftrightarrow M = \sum_{r=1}^{R} s_r \, u_r \circ v_r \circ w_r
\]

“CP” model (Harshman 1970, Carrol and Chang 1970)
Hoff 2011: Hierarchical Bayes approach
Generalizing the SVD, part 1

Mean model:

\[ Y = M + E \]

Matrices:

\[
\text{rank}(\mathbf{M} = r) \iff \mathbf{M} = \sum_{r=1}^{R} s_r \mathbf{u}_r \mathbf{v}_r^T = \sum_{r=1}^{R} s_r \mathbf{u}_r \circ \mathbf{v}_r = \mathbf{U} \mathbf{S} \mathbf{V}^T
\]

Arrays:

\[
\text{rank}(\mathbf{M} = r) \iff \mathbf{M} = \sum_{r=1}^{R} s_r \mathbf{u}_r \circ \mathbf{v}_r \circ \mathbf{w}_r
\]

“CP” model (Harshman 1970, Carrol and Chang 1970)

Hoff 2011: Hierarchical Bayes approach
Mean model:

\[ Y = M + E \]

Matrices:
- \( r_1 = \dim(\text{span(rows of } Y)) \)
- \( r_2 = \dim(\text{span(columns of } Y)) \)
- \( r_1 = r_2 \)

Arrays:
- \( r_k = \dim(\text{span(rows of } Y_{(k)})) \)
- It is possible that \( r_1 \neq r_2 \neq \cdots \neq r_K \)
Generalizing the SVD, part 2

Mean model:

\[ Y = M + E \]

Matrices:
- \( r_1 = \dim(\text{span(rows of } Y)) \)
- \( r_2 = \dim(\text{span(columns of } Y)) \)
- \( r_1 = r_2 \)

Arrays:
- \( r_k = \dim(\text{span(rows of } Y_{(k)})) \)
- It is possible that \( r_1 \neq r_2 \neq \cdots \neq r_K \).
Tensor SVD: If $M$ is of rank $r = (r_1, \ldots, r_K)$ then

$$M = S \times \{U_1, \ldots, U_K\}$$

- $S$ is the $r_1 \times \cdots \times r_K$ “core array”
- $U_k \in \mathbb{R}^{m_k \times r_k}$, $U_k^T U_k = I$.

(Delathauwer et al, Kolda and Bader)

Relation to matrix SVD:

$$M = U_1 S U_2^T \iff m = (U_2 \otimes U_1) s$$

$$M = S \times \{U_1, \ldots, U_K\} \iff m = (U_K \otimes \cdots \otimes U_1) s$$
**Tensor SVD**: If \( M \) is of rank \( r = (r_1, \ldots, r_K) \) then

\[
M = S \times \{U_1, \ldots, U_K\}
\]

- \( S \) is the \( r_1 \times \cdots \times r_K \) "core array"
- \( U_k \in \mathbb{R}^{m_k \times r_k}, U_k^T U_k = I \).

(Delathauwer et al, Kolda and Bader)

**Relation to matrix SVD**:

\[
M = U_1 S U_2^T \iff m = (U_2 \otimes U_1) s
\]

\[
M = S \times \{U_1, \ldots, U_K\} \iff m = (U_K \otimes \cdots \otimes U_1) s
\]
HOSVD

Tensor SVD: If $M$ is of rank $r = (r_1, \ldots, r_K)$ then

$$M = S \times \{U_1, \ldots, U_K\}$$

- $S$ is the $r_1 \times \cdots \times r_K$ “core array”
- $U_k \in \mathbb{R}^{m_k \times r_k}$, $U_k^T U_k = I$.

(Delathauwer et al, Kolda and Bader)

Relation to matrix SVD:

$$M = U_1 S U_2^T \iff m = (U_2 \otimes U_1) \ s$$

$$M = S \times \{U_1, \ldots, U_K\} \iff m = (U_K \otimes \cdots \otimes U_1) \ s$$
Tensor SVD model

\[ Y = \sigma(S \times \{U_1, \ldots, U_K\} + E) \]
\[ y = \sigma((U_K \otimes \cdots \otimes U_1) s + e) \]

This model is invariant under transformations of the form

\[ g : y \rightarrow \tau(W_K \otimes \cdots \otimes W_1) y = \tau Wy \]

This suggests we may want to use equivariant estimators:

\[ \hat{U}(\tau Wy) = W\hat{U}(y) \]
\[ \hat{\sigma}(\tau Wy) = \tau\hat{\sigma}(y) \]
Tensor SVD model

\[ \mathbf{Y} = \sigma(\mathbf{S} \times \{ \mathbf{U}_1, \ldots, \mathbf{U}_K \} + \mathbf{E}) \]
\[ \mathbf{y} = \sigma((\mathbf{U}_K \otimes \cdots \otimes \mathbf{U}_1) \mathbf{s} + \mathbf{e}) \]

This model is invariant under transformations of the form

\[ g : \mathbf{y} \rightarrow \tau(\mathbf{W}_K \otimes \cdots \otimes \mathbf{W}_1) \mathbf{y} = \tau \mathbf{W} \mathbf{y} \]

This suggests we may want to use equivariant estimators:

\[ \hat{\mathbf{U}}(\tau \mathbf{W} \mathbf{y}) = \mathbf{W} \hat{\mathbf{U}}(\mathbf{y}) \]
\[ \hat{\sigma}(\tau \mathbf{W} \mathbf{y}) = \tau \hat{\sigma}(\mathbf{y}) \]
Tensor SVD model

\[ Y = \sigma(S \times \{U_1, \ldots, U_K\} + E) \]
\[ y = \sigma((U_K \otimes \cdots \otimes U_1) s + e) \]

This model is invariant under transformations of the form

\[ g : y \rightarrow \tau(W_K \otimes \cdots \otimes W_1) y = \tau Wy \]

This suggests we may want to use equivariant estimators:

\[ \hat{U}(\tau Wy) = W\hat{U}(y) \]
\[ \hat{\sigma}(\tau Wy) = \tau\hat{\sigma}(y) \]
Equivariant estimation

If \( s \) were known, UMRE estimators are Bayes estimators under

- \( \pi(U_k) \) uniform density on \((\mathcal{N}_{r_k}, m_k)\)
- \( \pi(\sigma) \propto 1/\sigma \).

Unfortunately \( s \) is not generally known. Two candidate priors are

\[
s \sim N(0, \tau^2 I) \\
\hat{s} \sim N(0, \tau^2 \Lambda_k \otimes \cdots \otimes \Lambda_1)
\]

where \( \Lambda_k \) is diagonal with \( \text{tr}(\Lambda_k) = 1 \).

**Idea:** \( \Lambda_k \) can penalize the mode \( k \) rank \( r_k \):

\[
\text{Cov}[S_{(1)}] = c\Lambda_k = c
\begin{pmatrix}
\lambda_{k1} & 0 & \cdots & 0 \\
0 & \lambda_{k2} & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & \lambda_{kr_k}
\end{pmatrix}
\]

We will use a uniform prior on the simplex for each \( \Lambda_k \).

Stronger penalties can be obtained from other Dirichlet distributions.
Equivariant estimation

If $s$ were known, UMRE estimators are Bayes estimators under

- $\pi(U_k)$ uniform density on $(\mathcal{V}_{r_k,m_k})$
- $\pi(\sigma) \propto 1/\sigma$.

Unfortunately $s$ is not generally known. Two candidate priors are

$$s \sim N(0, \tau^2 I)$$
$$s \sim N(0, \tau^2 \Lambda_K \otimes \cdots \otimes \Lambda_1)$$

where $\Lambda_k$ is diagonal with $\text{tr}(\Lambda_k) = 1$.

**Idea:** $\Lambda_k$ can penalize the mode $k$ rank $r_k$:

$$\text{Cov}[S_{(1)}] = c\Lambda_k = c \begin{pmatrix}
\lambda_{k1} & 0 & \cdots & 0 \\
0 & \lambda_{k2} & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & \lambda_{kr_k}
\end{pmatrix}$$

We will use a uniform prior on the simplex for each $\Lambda_k$.

Stronger penalties can be obtained from other Dirichlet distributions.
Equivariant estimation

If $s$ were known, UMRE estimators are Bayes estimators under

- $\pi(U_k)$ uniform density on $(V_{r_k},m_k)$
- $\pi(\sigma) \propto 1/\sigma$.

Unfortunately $s$ is not generally known. Two candidate priors are

$$s \sim N(0, \tau^2 I)$$
$$s \sim N(0, \tau^2 \Lambda_K \otimes \cdots \otimes \Lambda_1)$$

where $\Lambda_k$ is diagonal with $\text{tr}(\Lambda_k) = 1$.

**Idea:** $\Lambda_k$ can penalize the mode $k$ rank $r_k$:

$$\text{Cov}[S_{(1)}] = c\Lambda_k = c\begin{pmatrix}
\lambda_{k1} & 0 & \cdots & 0 \\
0 & \lambda_{k2} & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & \lambda_{kr_k}
\end{pmatrix}$$

We will use a uniform prior on the simplex for each $\Lambda_k$.

Stronger penalties can be obtained from other Dirichlet distributions.
Equivariant estimation

If $s$ were known, UMRE estimators are Bayes estimators under

- $\pi(U_k)$ uniform density on $(\mathcal{V}_{r_k,m_k})$
- $\pi(\sigma) \propto 1/\sigma$.

Unfortunately $s$ is not generally known. Two candidate priors are

$$s \sim N(0, \tau^2 I)$$
$$s \sim N(0, \tau^2 \Lambda_K \otimes \cdots \otimes \Lambda_1)$$

where $\Lambda_k$ is diagonal with $\text{tr}(\Lambda_k) = 1$.

**Idea:** $\Lambda_k$ can penalize the mode $k$ rank $r_k$:

$$\text{Cov}[S^{(1)}] = c\Lambda_k = c \begin{pmatrix}
\lambda_{k1} & 0 & \cdots & 0 \\
0 & \lambda_{k2} & \cdots & 0 \\
\vdots & \ddots & \ddots & \vdots \\
0 & 0 & \cdots & \lambda_{kr_k}
\end{pmatrix}$$

We will use a uniform prior on the simplex for each $\Lambda_k$.

Stronger penalties can be obtained from other Dirichlet distributions.
Equivariant estimation

If \( s \) were known, UMRE estimators are Bayes estimators under

- \( \pi(U_k) \) uniform density on \((V_{r_k}, m_k)\)
- \( \pi(\sigma) \propto 1/\sigma \).

Unfortunately \( s \) is not generally known. Two candidate priors are

\[
s \sim N(0, \tau^2 I)
\]
\[
s \sim N(0, \tau^2 \Lambda_K \otimes \cdots \otimes \Lambda_1)
\]

where \( \Lambda_k \) is diagonal with \( \text{tr}(\Lambda_k) = 1 \).

**Idea:** \( \Lambda_k \) can penalize the mode \( k \) rank \( r_k \):

\[
\text{Cov}[S(1)] = c\Lambda_k = c
\begin{pmatrix}
\lambda_{k1} & 0 & \cdots & 0 \\
0 & \lambda_{k2} & \cdots & 0 \\
\vdots & \ddots & \ddots & \vdots \\
0 & 0 & \cdots & \lambda_{kr_k}
\end{pmatrix}
\]

We will use a uniform prior on the simplex for each \( \Lambda_k \).

Stronger penalties can be obtained from other Dirichlet distributions.
Simulation study

Consider estimation of $M$ with true rank vector $r_0$ under the model where

- $r = r_0$
- $r = 2 \times r_0$

We examined this under the following scenario

- $m = (60, 50, 40)$;
- $r_0 = (6, 5, 4)$, $r_0 = (30, 25, 20)$;
- two signal-to-noise ratios.
Consider estimation of $\mathbf{M}$ with true rank vector $\mathbf{r}_0$ under the model where

- $\mathbf{r} = \mathbf{r}_0$
- $\mathbf{r} = 2 \times \mathbf{r}_0$

We examined this under the following scenario

- $\mathbf{m} = (60, 50, 40)$;
- $\mathbf{r}_0 = (6, 5, 4)$, $r_0 = (30, 25, 20)$;
- two signal-to-noise ratios.
Simulation results

![Graph showing simulation results](image-url)
Extension to non-normal ordinal data

**Application**: $Y = \{y_{i,j,k,l}\}$ records relationships between countries
- $i = 1, \ldots, 50$ indexes actor countries;
- $j = 1, \ldots, 50$ indexes target countries;
- $k = 1, \ldots, 37$ indexes weeks in 2010;
- $l = 1, \ldots, 10$ indexes types of actions.
Model

\[ \mathbf{Z} = \mathbf{S} \times \{ \mathbf{U}_1, \ldots, \mathbf{U}_4 \} + \mathbf{E} \]

\[ y_{i,j,k,l} = f_i(z_{i,j,k,l}) \]

\( f_1, \ldots, f_{m_K} \) are unknown and arguably a nuisance parameter.

Estimation for \( \mathbf{U} \) and \( \mathbf{S} \) can proceed via the rank likelihood (Hoff, 2007).
Model

\[ Z = S \times \{ U_1, \ldots, U_4 \} + E \]

\[ y_{i,j,k,l} = f_i(z_{i,j,k,l}) \]

\( f_1, \ldots, f_{m_K} \) are unknown and arguably a nuisance parameter.

Estimation for \( U \) and \( S \) can proceed via the rank likelihood (Hoff, 2007).
Estimation results
Estimation results
Summary and comments

Summary:
1. multi-indexed data can be represented by arrays
   • data or parameters can be an array
2. array representations force consideration of heterogeneity along each mode
3. mean and covariance models available via multilinear products

Comments:
1. There exist a great variety of factor models and decompositions. How to choose?
   • model selection via hypothesis testing?
   • give up on interpretability - fit a large model and penalize.
2. Need to move beyond least squares.
Summary and comments

Summary:
1. multi-indexed data can be represented by arrays
   - data or parameters can be an array
2. array representations force consideration of heterogeneity along each mode
3. mean and covariance models available via multilinear products

Comments:
1. There exist a great variety of factor models and decompositions. How to choose?
   - model selection via hypothesis testing?
   - give up on interpretability - fit a large model and penalize.
2. Need to move beyond least squares.