Bayesian Regression with Input Noise for High Dimensional Data

Jo-Anne Ting[1], Aaron D'Souza[2], Stefan Schaal[1]

[1]University of Southern California,

[2]Google, Inc.

June 26, 2006

Agenda

- Relevance of high dimensional regression with input noise

- Introduction to Bayesian parameter estimation
  - EM-based Joint Factor Analysis
  - Automatic feature detection
  - Making predictions with noiseless query points

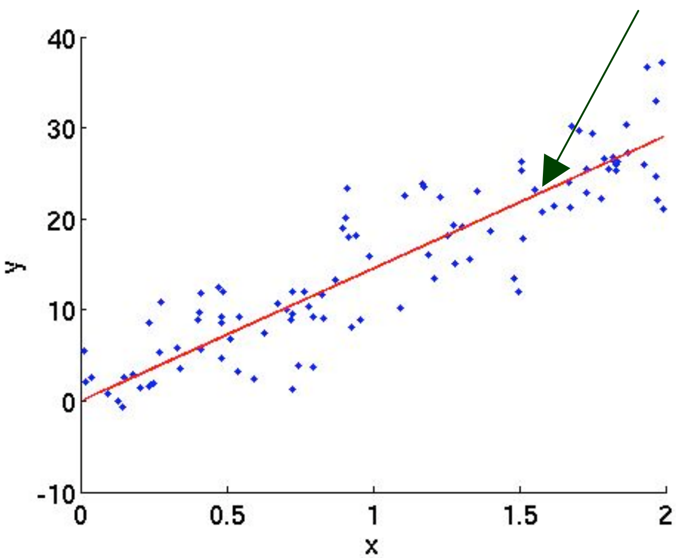
- Evaluation on a 100-dimensional synthetic dataset

- Application to Rigid Body Dynamics parameter identification
  - What are RBD parameters?
  - Formulate it as a linear regression problem
  - How to ensure physically consistent parameters?

We are interested in parameter estimation…

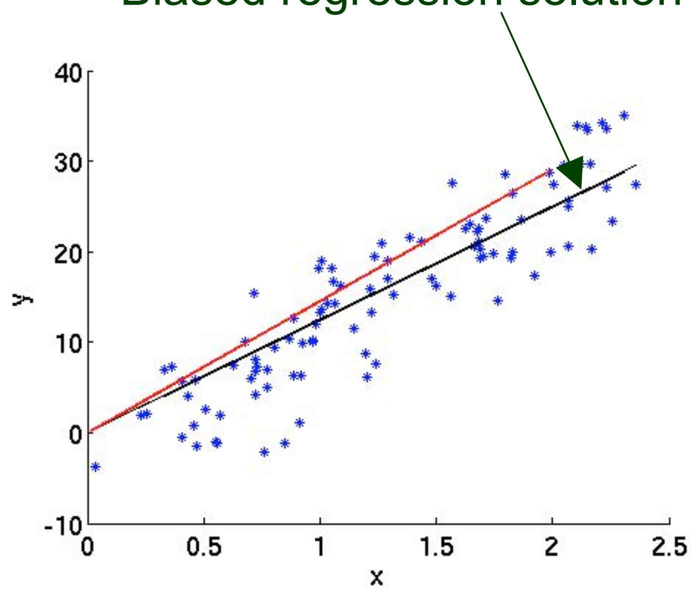
- Traditional regression techniques ignore noise in input data, for example, for linear regression*:

Unbiased regression solution



Noiseless inputs

Biased regression solution



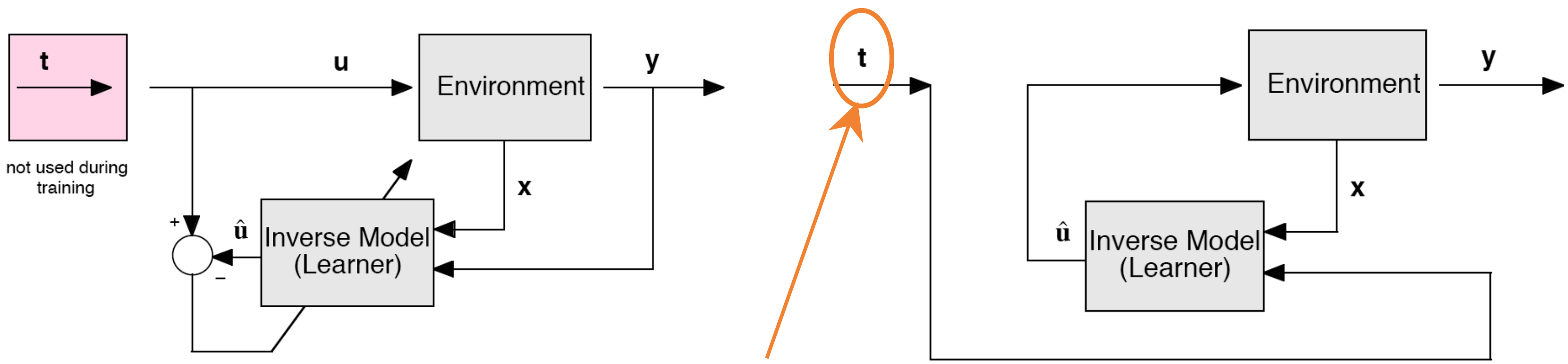
Noisy inputs

* Solutions to linear problems can be easily extended to nonlinear systems via locally weighted methods (e.g. Atkeson et al. 1997)

…and Prediction With Noiseless Query Points

- For physical systems such as humanoid robots:
  - Noisy input data, large number of input dimensions -- of which not all is relevant

- We want to control these robots using model-based controllers:



Training Phase

t is the desired (noiseless) target

Testing Phase

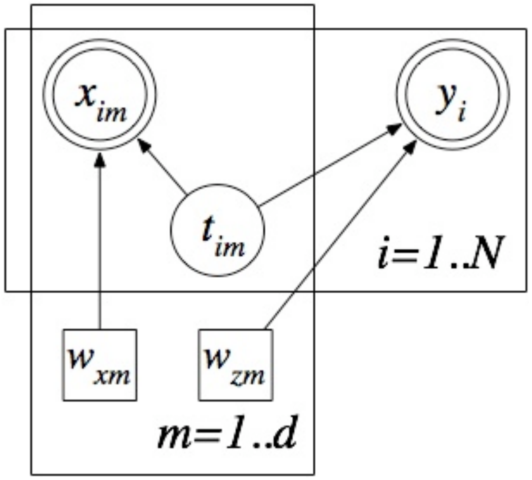
Current Methods Are Unsuitable

	Ignores input noise	*Accounts for input noise*
Unsuitable for high dimensional data	• OLS with robust matrix inversion (e.g. Belsley et al. 1980): $O(d^2)$ at best	• Total LS/Orthogonal LS (e.g. Golub & VanLoan 1998, Hollerbach & Wampler 1996) • Joint Factor Analysis (JFA) (Massey 1965): computationally prohibitive in high dimensions
Suitable for high dimensional data	• LASSO & Stepwise regression (Tibshirani 1996, Draper & Smith 1981)	???

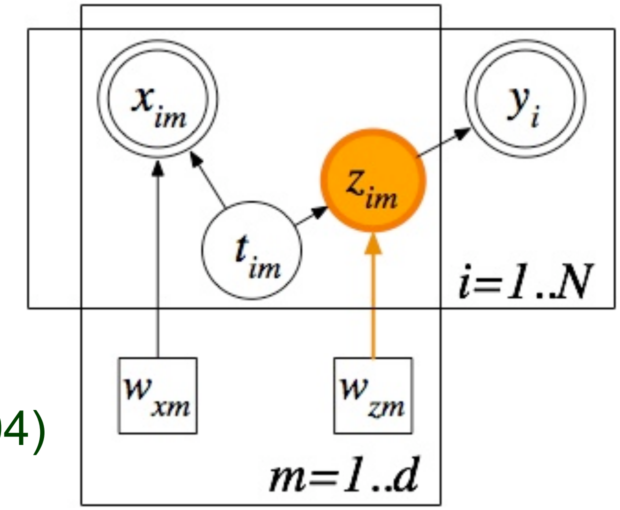
Agenda

- Relevance of high dimensional regression with input noise
- Introduction to Bayesian parameter estimation
  - EM-based Joint Factor Analysis
  - Automatic feature detection
  - Making predictions with noiseless query points
- Evaluation on a 100-dimensional synthetic dataset
- Application to Rigid Body Dynamics parameter identification
  - What are RBD parameters?
  - Formulate it as a linear regression problem
  - How to ensure physically consistent parameters?

Computationally Prohibitive? Not Any More!



Introduce hidden variables, z_{im} (D'Souza et al. 2004)



$$y_i = \sum_{m=1}^{d} w_{zm} t_{im} + \varepsilon_y$$

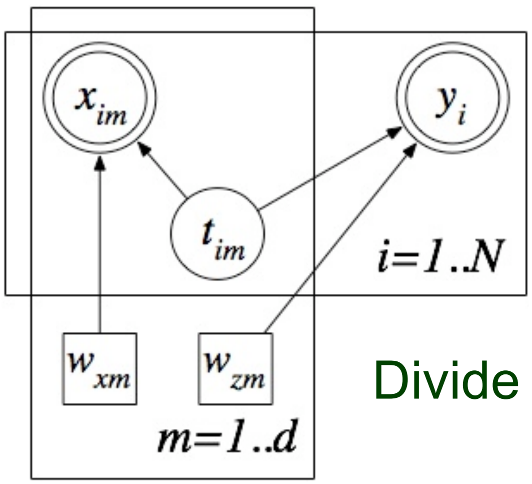
$$x_i = \sum_{m=1}^{d} w_{xm} t_{im} + \varepsilon_x$$

$$y_i = \sum_{m=1}^{d} z_{im} + \varepsilon_y$$

$$z_{im} = \sum_{m=1}^{d} w_{zm} t_{im} + \eta_m$$

EM-based JFA: All EM update equations are O(d)

…but Remember the Important Parameters



JFA

$$y_i - \varepsilon_y = \sum_{m=1}^{d} w_{zm} t_{im} \quad (1)$$

$$x_i - \varepsilon_x = \sum_{m=1}^{d} w_{xm} t_{im} \quad (2)$$

Divide (1) by (2) to get:

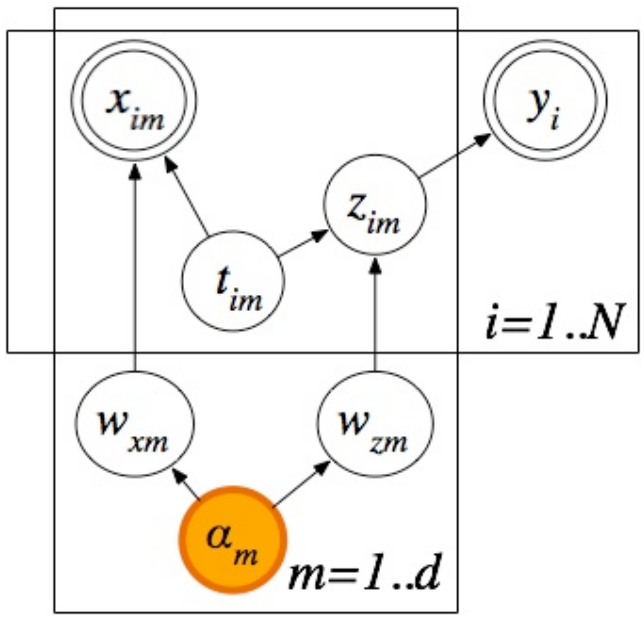
$$\frac{y_i - \varepsilon_y}{x_i - \varepsilon_x} = \frac{\sum_{m=1}^{d} w_{zm} t_{im}}{\sum_{m=1}^{d} w_{xm} t_{im}}$$

or...

$$y_i = \sum_{m=1}^{d} \frac{w_{zm}}{w_{xm}} \left(x_i - \varepsilon_x\right) + \varepsilon_y$$

This is the solution to the regression problem $y=b^T x$ -- which is what we need for prediction

Next, We Add Automatic Feature Detection



Priors:

$$p(\alpha_m) = \text{Gamma}(a_m, b_m)$$

$$p(w_{zm} | \alpha_m) = \text{Normal}\left(0, {}^{1}\!/_{\alpha_m}\right)$$

$$p(w_{xm} | \alpha_m) = \text{Normal}\left(0, {}^{1}\!/_{\alpha_m}\right)$$

- Coupled regularization of regression parameters
- Still O(d) per EM iteration

Making Predictions with Noiseless Query Points

- For a noisy test input x^q and its unknown output y^q, $\hat{b}_{noise} = ?$

Given:

$$p\left(y^q \middle| \mathbf{x}^q\right) = \int\int p\left(y^q, \mathbf{Z}, \mathbf{T} \middle| \mathbf{x}^q\right) d\mathbf{Z} d\mathbf{T}$$

$$\left\langle y^q \middle| \mathbf{x}^q \right\rangle = \hat{b}_{noise}^T x^q$$

We can infer:

$$\hat{b}_{noise} = \frac{\psi_y 1^T \mathbf{B}^{-1}}{\psi_y - 1^T \mathbf{B} 1} \boldsymbol{\Psi}_z^{-1} \langle \mathbf{W}_z \rangle \mathbf{A}^{-1} \langle \mathbf{W}_x \rangle^T \boldsymbol{\Psi}_x^{-1}$$

- For a noiseless test input t^q and its unknown output y^q, $\hat{b}_{true} = ?$

$$\lim_{\psi_x \to 0} \hat{b}_{noise} \Leftrightarrow \hat{b}_{true} = \frac{\psi_y 1^T \mathbf{C}^{-1}}{\psi_y - 1^T \mathbf{C}^{-1} 1} \boldsymbol{\Psi}_z^{-1} \langle \mathbf{W}_z \rangle \langle \mathbf{W}_x \rangle^{-1}$$

...where $\mathbf{C} = \left(11^T \big/ \psi_y + \boldsymbol{\Psi}_z^{-1} \right)$

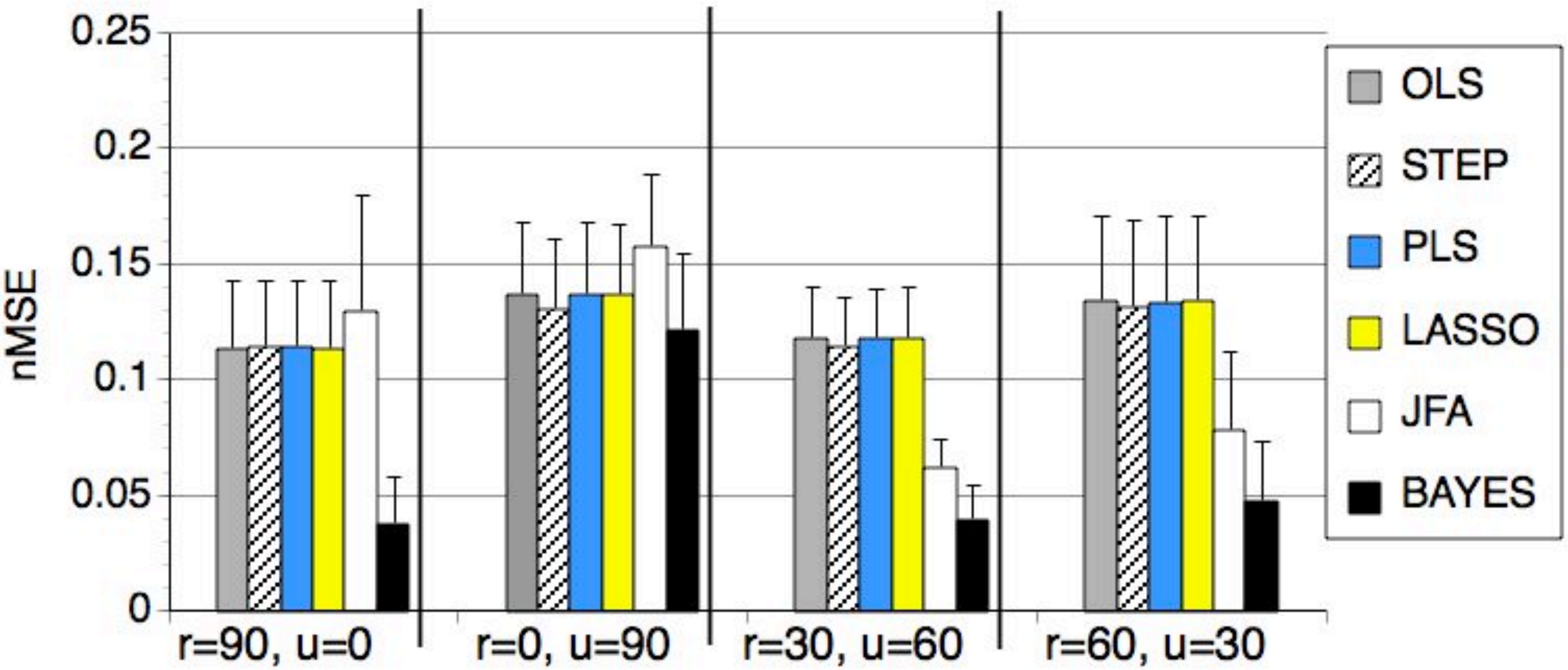
Agenda

- Relevance of high dimensional regression with input noise
- Introduction to Bayesian parameter estimation
  - EM-based Joint Factor Analysis
  - Automatic feature detection
  - Making predictions using noiseless query points
- Evaluation on a 100-dimensional synthetic dataset
- Application to Rigid Body Dynamics parameter identification
  - What are RBD parameters?
  - Formulate it as a linear regression problem
  - How to ensure physically consistent parameters?

Construction of 100-dimensional dataset

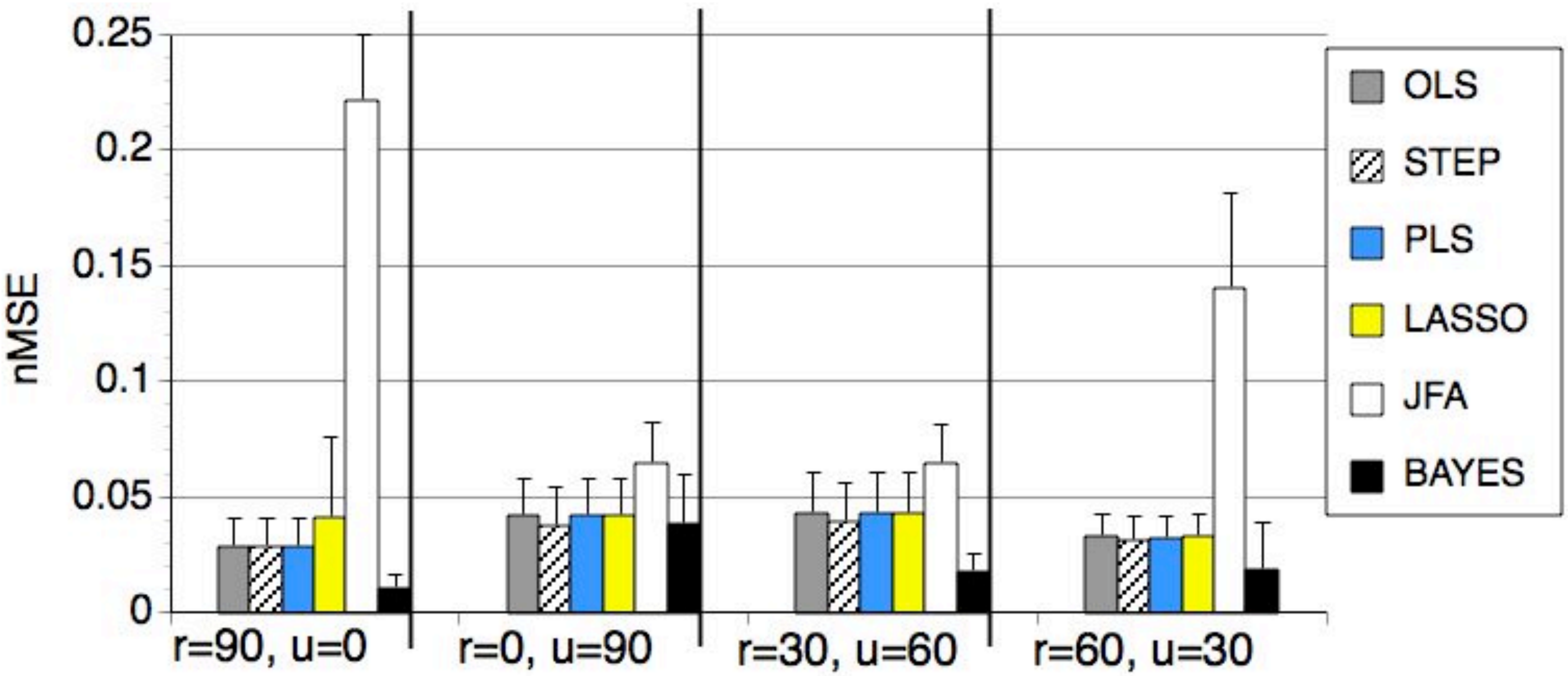
- Constructed data with
  - 10 relevant dimensions
  - 90 redundant and/or irrelevant dimensions
- Explored different combinations of redundant (r) and irrelevant (u) dimensions
  - r = 90, u = 0: 90 redundant dimensions
  - r = 0, u = 90: 90 irrelevant dimensions
  - r = 30, u = 60
  - r = 60, u = 30
- Tested on strongly noisy (SNR=2) and less noisy (SNR=5) data
- Predicted outputs with noiseless test inputs

10-70% Improvement for Strongly Noisy Data (SNR=2)



Bayesian parameter estimation generalizes *10-70%* better for strongly noisy data

7-50% Improvement on Less Noisy Data (SNR=5)



…and *7-50%* better for less noisy data

Agenda

- Relevance of high dimensional regression with input noise
- Introduction to Bayesian parameter estimation
  - EM-based Joint Factor Analysis
  - Automatic feature detection
  - Making predictions with noiseless query points
- Evaluation on a 100-dimensional synthetic dataset
- Application to Rigid Body Dynamics parameter identification
  - What are RBD parameters?
  - Formulate it as a linear regression problem
  - How to ensure physically consistent parameters?

What are Rigid Body Dynamics (RBD) Parameters?

- Using Newton-Euler equations for a rigid body, we get the RBD equation (where *q* are joint angles):

$$\tau = M(q)\ddot{q} + C(\dot{q}, q) + G(q)$$

Mass matrix — Centripetal & Coriolis terms — Vector of gravity terms

- *M*, *C* and *G* are functions of mass, centre of mass and moments of inertia -- all which are unknown; *q*'s and τ are known

- We can re-express the above linearly as:

$$\tau = Y(q, \dot{q}, \ddot{q})\theta$$

Formulate RBD Parameter Identification As A Linear Regression Problem

(e.g. An et al. 1988)

$$\tau = Y\left(q, \dot{q}, \ddot{q}\right)\theta$$

where the RBD parameters are…

$$\theta = [m,\ mc_x,\ mc_y,\ mc_z,\ I_{11},\ I_{12},\ I_{13},\ I_{22},\ I_{23},\ I_{33}]^T$$

- RBD parameters:
  - Must satisfy physical constraints (positive mass, positive definite inertia matrix)
  - But.. not all parameters are identifiable due to insufficiently rich data & constraints of the physical system (i.e. data is ill-conditioned)

Specifically, a High Dimensional Noisy Linear Regression Problem

- To enforce physical constraints on θ, introduce virtual parameters $\hat{\theta}$:

$$\theta_1 = \hat{\theta}_1^2,\ \theta_2 = \hat{\theta}_2\hat{\theta}_1^2,\ \theta_3 = \hat{\theta}_3\hat{\theta}_1^2$$

$$\theta_4 = \hat{\theta}_4\hat{\theta}_1^2,\ \theta_5 = \hat{\theta}_5^2 + \left(\hat{\theta}_4^2 + \hat{\theta}_3^2\right)\hat{\theta}_1^2$$

$$\theta_6 = \hat{\theta}_5\hat{\theta}_6 - \hat{\theta}_2\hat{\theta}_3\hat{\theta}_1^2,\ \theta_7 = \hat{\theta}_5\hat{\theta}_7 - \hat{\theta}_2\hat{\theta}_4\hat{\theta}_1^2$$

$$\theta_8 = \hat{\theta}_6^2 + \hat{\theta}_8^2 + \left(\hat{\theta}_2^2 + \hat{\theta}_4^2\right)\hat{\theta}_1^2$$

$$\theta_9 = \hat{\theta}_6\hat{\theta}_7 + \hat{\theta}_8\hat{\theta}_9 - \hat{\theta}_3\hat{\theta}_4\hat{\theta}_1^2$$

$$\theta_{10} = \hat{\theta}_7^2 + \hat{\theta}_9^2 + \hat{\theta}_{10}^2 + \left(\hat{\theta}_2^2 + \hat{\theta}_3^2\right)\hat{\theta}_1^2,\ \theta_{11} = \hat{\theta}_{11}^2$$

- 11 features per DOF
- For a system with *s* DOF, there are 11*s* features

- Consequently, for real world systems, we have a noisy, high dimensional, ill-conditioned linear regression problem

How to Ensure Our Robust Parameter Estimates are Physically Consistent?

- Find physically consistent robust parameter estimates that are as close to $\hat{b}_{true}$ as possible

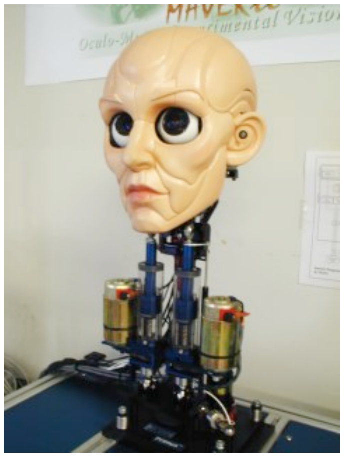
- Do a constraint optimization step to find $\hat{\theta}_{optimal}$:

$$\hat{\theta}_{optimal} = \arg\min_{\hat{\theta}} w\left[\hat{b}_{true} - f\left(\hat{\theta}\right)\right]$$

where $w_m = 0$ if dimension m is not relevant and $w_m = 1$ otherwise

- Finally, ensure redundant/irrelevant dimensions in $\hat{b}_{true}$ remain so in $\theta_{optimal}$

10-20% Improvement on Robotic Oculomotor Vision Head



- 7 DOFs: 3 in neck, 2 in each eye
- 11 features per DOF; total of *77 features*
- RBD parameter estimates from *ALL* algorithms satisfy physical constraints
- Bayesian de-noising does *~10-20% better*

Root Mean Squared Errors

Algorithm	Position(rad)	Velocity(rad/s)	Feedback (Nm)
Ridge regression	0.0291	0.2465	0.3969
Bayesian de-noising	0.0243	0.2189	0.3292
LASSO regression	0.0308	0.2517	0.4274
Stepwise regression	FAILURE	FAILURE	FAILURE

5-17% Improvement on Robotic Anthropomorphic Arm

- 10 DOFs: 3 in shoulder, 1 in elbow, 3 in wrist, 3 in fingers
- 11 features per DOF; total of *110 features*
- Bayesian de-noising does *~5-17% better*

Root Mean Squared Errors

Algorithm	Position(rad)	Velocity(rad/s)	Feedback (Nm)
Ridge regression	0.0210	0.1119	0.5839
Bayesian de-noising	0.0201	0.0930	0.5297
LASSO regression	FAILURE	FAILURE	FAILURE
Stepwise regression	FAILURE	FAILURE	FAILURE

Summary

- Bayesian treatment of Joint Factor Analysis that performs parameter estimation with noisy input data
- O(d) complexity per EM iteration
- Automatic feature detection through joint regularization of both regression branches
- Significant improvement on synthetic data and real-world systems