Lecture Notes on

Info-Gap Estimation and Forecasting

Yakov Ben-Haim
Yitzhak Moda'i Chair in
Technology and Economics
Faculty of Mechanical Engineering
Technion — Israel Institute of Technology
Haifa 32000 Israel

http://info-gap.com Blog: http://decisions-and-info-gaps.blogspot.com http://www.technion.ac.il/yakov
Tel: +972-4-829-3262, +972-50-750-1402 Fax: +972-4-829-5711
yakov@technion.ac.il

¶ Source material:

• Yakov Ben-Haim, 2005, Info-gap Decision Theory For Engineering Design. Or: Why 'Good' is Preferable to 'Best', appearing as chapter 11 in *Engineering Design Reliability Handbook,* Edited by Efstratios Nikolaidis, Dan M.Ghiocel and Surendra Singhal, CRC Press, Boca Raton.

• Yakov Ben-Haim, 2006, *Info-Gap Decision Theory: Decisions Under Severe Uncertainty,* 2nd edition, section 3.2.13, Academic Press, London.

• Yakov Ben-Haim, 2010, *Info-Gap Economics: An Operational Introduction,* Palgrave-Macmillan, London.

• Yakov Ben-Haim, 2018, *Dilemmas of* Wonderland: *Decisions in the Age of Innovation,* Oxford University Press.

• Yakov Ben-Haim, 2008, Info-gap forecasting and the advantage of sub-optimal models, *European Journal of Operational Research,* 197: 203–213.

• Yakov Ben-Haim, 2008, Info-Gap Economics: An Overview, working paper. (\papers\BoE2008\ige03.tex)

A Note to the Student: These lecture notes are not a substitute for the thorough study of books. These notes are no more than an aid in following the lectures.

Contents

1 Linear Regression

1.1 Preliminary Discussion

¶ **Modeling is a decision problem.** We will consider 3 examples:
- Modeling WLAN client position and predicting next location.
- Modeling a mechanical S-N curve.
- Modeling the economic Phillips curve.[1]

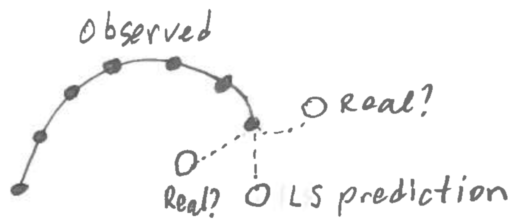


Figure 1: WLAN client motion.

¶ **WLAN client tracking and prediction:**
¶ Challenge: Two foci of uncertainty:
- Randomness:
  - Noisy data (statistics).
- Info-gaps:
  - Changing plans and intentions of client.
  - Interaction with other people.
  - Environmental variability.

¶ Questions:
- How to use empirical data to model uncertain past motion?
- Is optimal estimation (e.g. least-squares) a good strategy for predicting future position?
- Can we do better?
- How to manage both statistical and info-gap uncertainty?
- How to evaluate estimates vis a vis info-gaps?

[1] Source: Yakov Ben-Haim, 2010, *Info-Gap Economics: An Operational Introduction,* Palgrave-Macmillan.

¶ Mechanical S-N curve:

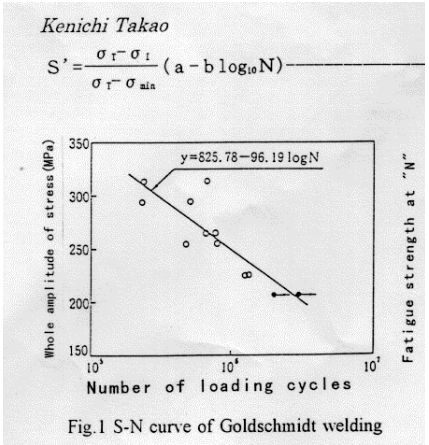


Figure 2: S-N curves.

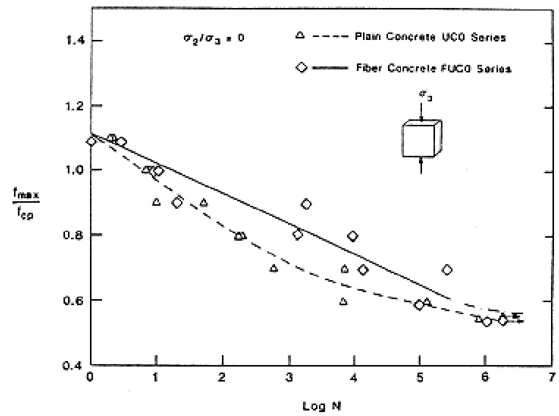


Figure 3: S-N curves.

¶ Challenge: Two foci of uncertainty:
- Randomness:
  - ○ Noisy data (statistics).
- Info-gaps:
  - ○ Changing fundamentals.
  - ○ Material variability.
  - ○ Environmental variability.

¶ Questions:
- How to use empirical data to model uncertain material?
- Is optimal estimation (e.g. least-squares) a good strategy?
- Can we do better?
- How to manage both statistical and info-gap uncertainty?
- How to evaluate estimates vis a vis info-gaps?

¶ **Economic Phillips curve:**

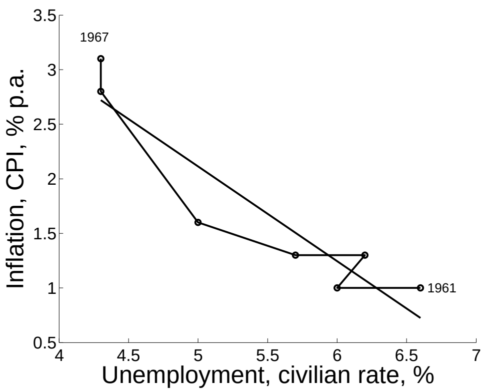


Figure 4: Inflation vs. unemployment in the US, 1961–1967.

Figure 5: Inflation vs. unemployment in the US, 1961–1993.

¶ Inflation vs. unemployment, US, '61–'67:
- Approximately linear.
- Slope ≈ -0.87 %CPI/%unemployment.

¶ Slopes in other periods:
- '61–'67: -0.87 • '80–'83: -3.34 • '85–'93: -1.08 • '70–'78: ???

¶ Challenge: Two foci of uncertainty:
- Randomness:
  - ◦ Noisy data (statistics).
- Info-gaps:
  - ◦ Changing fundamentals.
  - ◦ Data revision.

¶ Questions:
- How to use historical data to model the future?
- Is optimal estimation (e.g. least-squares) a good strategy?
- Can we do better?
- How to manage both statistical and info-gap uncertainty?
- How to evaluate estimates vis a vis info-gaps?

1.2 Robustness with Fractional-Error Parameter Uncertainty

1.2.1 Formulation of the Problem

¶ **Paired data,** fig. 6:
- System lifetime, CPI, etc: $c_1, \ldots, c_n$.
- Mechanical stress, unemployment, etc: $u_1, \ldots, u_n$.

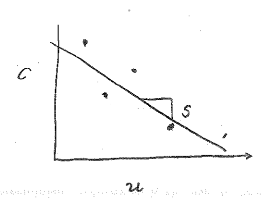


Figure 6: Paired data.

¶ **Least-squares estimate of slope, e.g. Young's modulus of stress-strain curve:**
- Linear regression:

$$c = su + b \tag{1}$$

- Mean squared error (MSE):

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}[c_i - (su_i + b)]^2 \tag{2}$$

- MSE estimate of the slope:

$$\widetilde{s} = \arg\min_{s} \text{MSE} \tag{3}$$

One finds:

$$\widetilde{s} = \frac{\operatorname{cov}(u,c)}{\operatorname{var}(u)} \tag{4}$$

where:

$$\operatorname{cov}(u,c) = \frac{1}{n}\sum_{i=1}^{n} c_i u_i - \left(\frac{1}{n}\sum_{i=1}^{n} c_i\right)\left(\frac{1}{n}\sum_{i=1}^{n} u_i\right) \tag{5}$$

and $\operatorname{var}(u) = \operatorname{cov}(u,u)$.
- In our case, fig. 6, $\widetilde{s} < 0$.

1.2.2 Formulation of the Robustness Function

¶ **Robustness question:**
- How much can the data err due to info-gaps, and the slope's error will be acceptable?
- Recall two types of errors: statistical and info-gap. We focus on info-gaps.

¶ **Notation for Moments:**
γ = covariance, $\operatorname{cov}(u,c)$. $\widetilde{\gamma}$ = estimate.
σ^2 = variance, $\operatorname{var}(u)$. $\widetilde{\sigma}^2$ = estimate.

¶ **Consider info-gaps in data.** Specifically, unknown fractional errors of moments:

$$\left|\frac{\gamma-\widetilde{\gamma}}{\widetilde{\gamma}}\right|, \quad \left|\frac{\sigma^2-\widetilde{\sigma}^2}{\widetilde{\sigma}^2}\right| \tag{6}$$

¶ **Fractional-error info-gap model of uncertainty:**

$$\mathcal{U}(h)=\left\{(\gamma,\sigma^2):\ \left|\frac{\gamma-\widetilde{\gamma}}{\widetilde{\gamma}}\right|\le h,\ \left|\frac{\sigma^2-\widetilde{\sigma}^2}{\widetilde{\sigma}^2}\right|\le h,\ \sigma^2\ge 0\right\},\quad h\ge 0$$

¶ **Least-squares estimate:** $\widetilde{s}=\widetilde{\gamma}/\widetilde{\sigma}^2$.
Actual value: $s=\gamma/\sigma^2$.

¶ **Satisficing performance requirement:** $|s(\gamma,\sigma^2)-\widetilde{s}|\le r_{\rm c}$.

¶ **Robustness of LS estimate $\widetilde{s}$:**
- Maximum tolerable uncertainty.
- Max horizon of uncertainty in moments at which $\widetilde{s}$ errs no more than $r_{\rm c}$:

$$\widehat{h}(\widetilde{s},r_{\rm c})=\max\left\{h:\ \left(\max_{\gamma,\sigma^2\in\mathcal{U}(h)}|s(\gamma,\sigma^2)-\widetilde{s}|\right)\le r_{\rm c}\right\} \tag{7}$$

1.2.3 Derivation of the Robustness Function

¶ **Derivation of the robustness:**
- $m(h)=$ inner maximum in eq.(7).
- $m(h)$ occurs at $\gamma=(1+h)\widetilde{\gamma},\ \sigma^2=(1-h)^+\widetilde{\sigma}^2$.
- Thus, for $h\le 1$:

$$m(h) = \left|\frac{(1+h)\widetilde{\gamma}}{(1-h)\widetilde{\sigma}^2}-\frac{\widetilde{\gamma}}{\widetilde{\sigma}^2}\right| \tag{8}$$

$$= \left(\frac{1+h}{1-h}-1\right)\left|\frac{\widetilde{\gamma}}{\widetilde{\sigma}^2}\right| \tag{9}$$

$$= \frac{2h}{1-h}|\widetilde{s}| \tag{10}$$

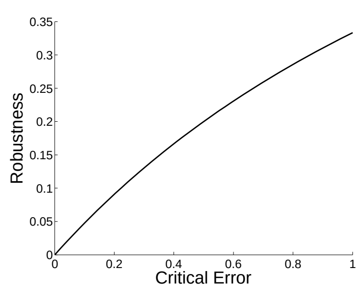


Figure 7: Robustness of estimated slope, $\widehat{h}(\widetilde{s},\rho)$, vs. critical error, ρ. Eq.(12).

- Equate $m(h) = r_{\rm c}$ and solve for h (recall $\widetilde{s} < 0$):

$$\frac{2h}{1-h} = -\frac{r_{\rm c}}{\widetilde{s}} = \rho \text{ (definition)} \quad \Longrightarrow \quad \widehat{h} = \frac{\rho}{2+\rho} \ \ (\leq 1) \tag{11}$$

¶ **Robustness of LS estimate $\widetilde{s}$:**

$$\widehat{h}(\widetilde{s}, \rho) = \frac{\rho}{2+\rho}, \quad \rho = -r_{\rm c}/\widetilde{s} \tag{12}$$

Recall: $\widetilde{s} < 0$ so $\rho > 0$.

- **Zeroing:** Relying on the best-estimate, $\widetilde{s}$, (so $r_{\rm c} = 0$), has zero robustness.
- **Trade-off:** robustness goes up (good) as required estimation error goes up (bad).
- **Examples:** $\rho = 0.0$, $\widehat{h} = 0$. $\rho = 0.2$, $\widehat{h} = 0.09$. See fig. 7, p.7.

1.2.4 Can We Do Better? Crossing Robustness Curves

¶ Can we do better than the LS estimate?

¶ Estimates of slope:

- $\widetilde{s}$ = LS estimate, with robustness $\widehat{h}(\widetilde{s}, r_{\rm c})$.
- $s_{\rm e}$= any estimate, with robustness $\widehat{h}(s_{\rm e}, r_{\rm c})$.
- Definitions: $\zeta = s_{\rm e}/\widetilde{s}, \quad \rho = -r_{\rm c}/\widetilde{s}$. (Recall: $\widetilde{s} < 0$.)
- Robustness of $s_{\rm e}$, in analogy to eq.(7):

$$\widehat{h}(s_{\rm e}, r_{\rm c}) = \max\left\{h: \left(\max_{\gamma,\sigma^2\in\mathcal{U}(h)} |s(\gamma,\sigma^2) - s_{\rm e}|\right) \le r_{\rm c}\right\} \tag{13}$$

◦ Let $m(h)$ denote the inner minimum:

$$m(h) = \max_{\gamma,\sigma^2\in\mathcal{U}(h)} \left|\frac{\gamma}{\sigma^2} - s_{\rm e}\right| \tag{14}$$

◦ For $h \le 1$ this occurs at one of the following:

$$\text{Either:} \quad \gamma = (1+h)\widetilde{\gamma}, \ \ \sigma^2 = (1-h)\widetilde{\sigma}^2 \tag{15}$$

$$\text{Or:} \quad \gamma = (1-h)\widetilde{\gamma}, \ \ \sigma^2 = (1+h)\widetilde{\sigma}^2 \tag{16}$$

◦ Denote the corresponding $m(h)$'s:

$$m_1(h) = \left|\frac{(1+h)\widetilde{\gamma}}{(1-h)\widetilde{\sigma}^2} - s_{\rm e}\right| \tag{17}$$

$$m_2(h) = \left|\frac{(1-h)\widetilde{\gamma}}{(1+h)\widetilde{\sigma}^2} - s_{\rm e}\right| \tag{18}$$

◦ $m(h)$ is the greater of these two alternatives:

$$m(h) = \max[m_1(h), \ m_2(h)] \tag{19}$$

The maximum depends on the value of h.

◦ After some algebra, recalling $\rho = -r_{\rm c}/\widetilde{s}$, and equating $m(h) = r_{\rm c}$, one finds:

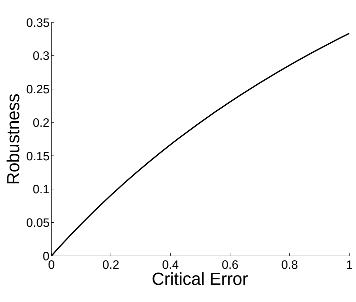


Figure 8: $\widehat{h}(\widetilde{s}, \rho)$ vs. ρ.

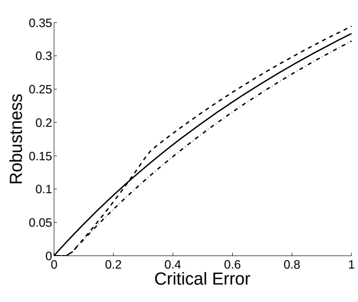


Figure 9: $\widehat{h}(s_{\rm e}, \rho)$ vs. ρ. $\zeta = 1$ (solid), 1.05 (dash), 0.95 (dot dash).

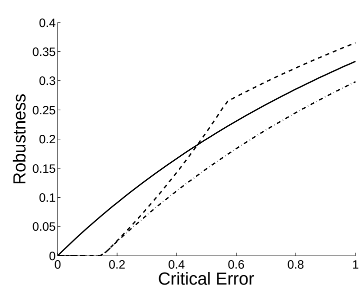


Figure 10: $\widehat{h}(s_{\rm e}, \rho)$ vs. ρ. $\zeta = 1$ (solid), 1.15 (dash), 0.85 (dot-dash).

$$\widehat{h}(s_{\rm e}, \rho) = \begin{cases} \dfrac{\rho+\zeta-1}{\rho+\zeta+1} & \text{if } \rho^2 \ge \zeta^2 - 1 \text{ and } \rho \ge 1-\zeta \\[2ex] \dfrac{\rho-\zeta+1}{-\rho+\zeta+1} & \text{if } \rho^2 \le \zeta^2 - 1 \text{ and } \rho \ge \zeta - 1 \end{cases} \tag{20}$$

$\widehat{h}(s_{\rm e}, \rho)$ is zero otherwise. Note $\widehat{h} \le 1$.

- Eq.(20) includes eq.(12) as a special case, when $\zeta = 1$.
- When $\zeta > 1$, the robustness follows the lower line of eq.(20) (which has greater slope than the robustness curve for $\widetilde{s}$) for small ρ, and then follows the upper line of the equation for larger ρ. This causes crossing of robustness curves as illustrated by the solid and dashed lines in figs. 9 and 10. (The two lines in eq.(20) are equal when $\rho^2 = \zeta^2 - 1$.)
- LS estimate: 0 error, 0 robustness.
- Trade-off: robustness vs. estim. error.
- Curve crossing: preference reversal.

¶ **Can we do better than least-squares?** Yes, but at a price:

- Robust-satisficing estimate is more robust to uncertainty at positive estimation error.
- Crossing robustness curves represents potential for **preference reversal.**

1.3 Robustness with Fourier-Ellipsoid Functional Uncertainty

1.3.1 Formulation of the Robustness Function

¶ Consider an uncertain function, $\sigma(\varepsilon)$, that may be stress as a function of strain, or any of the functional relations considered before, or any other functional relation.

¶ We represent $\sigma(\varepsilon)$, the **system model,** as:

$$\sigma(\varepsilon) = \widetilde{\sigma}(\varepsilon) + s(\varepsilon) \tag{21}$$

where:
$\widetilde{\sigma}(\varepsilon) =$ putative best estimate of the function.
$s(\varepsilon) =$ unknown error of the putative best estimate.

¶ What we know about the error function, $s(\varepsilon)$, is:
- The frequency range within which it can be represented as a truncated Fourier series.
- Estimates of the Fourier amplitudes.

That is, N is known such that:

$$s(\varepsilon) = \sum_{n=0}^{N} \beta_n \cos n\pi\varepsilon = \beta^T \gamma(\varepsilon) \tag{22}$$

The best estimate of the function is based on estimated Fourier coefficients $\widetilde{\beta}$:

$$\widetilde{s}(\varepsilon) = \sum_{n=0}^{N} \widetilde{\beta}_n \cos n\pi\varepsilon = \widetilde{\beta}^T \gamma(\varepsilon) \tag{23}$$

¶ The **uncertainty** in the Fourier coefficients is represented by this ellipsoid-bound info-gap model:

$$\mathcal{U}(h) = \left\{ \beta : \ \left(\beta - \widetilde{\beta}\right)^T W \left(\beta - \widetilde{\beta}\right) \le h^2 \right\}, \quad h \ge 0 \tag{24}$$

where W is a known, real, symmetric, positive definite matrix; something like an inverse covariance matrix.
- Note the zeroing and nesting properties.
- **What is the intuitive meaning of this info-gap model?**

¶ Our **performance requirement** is that the estimated stress, $\widetilde{\sigma}(\varepsilon)$, under-estimates the true stress, $\sigma(\varepsilon)$, by no more than δ:

$$\sigma(\varepsilon) - \widetilde{\sigma}(\varepsilon) \le \delta \tag{25}$$

¶ One could consider other requirements, such as a symmetric error requirement:

$$|\sigma(\varepsilon) - \widetilde{\sigma}(\varepsilon)| \le \delta \tag{26}$$

¶ The robustness to uncertainty in the stress function is the greatest horizon of uncertainty, h, up to which all Fourier coefficient vectors in the uncertainty set $\mathcal{U}(h)$ of eq.(24) satisfy the performance requirement in eq.(25):

$$\widehat{h}(\delta) = \max\left\{h:\ \left(\max_{\beta\in\mathcal{U}(h)}\left[\sigma(\varepsilon)-\widetilde{\sigma}(\varepsilon)\right]\right)\le\delta\right\} \tag{27}$$

¶ Locate the 3 components of the robustness function in eq.(27):
system model, uncertainty model, performance requirement?

1.3.2 Derivation of the Robustness function

¶ Let $m(h)$ denote the inner maximum in the definition of the robustness function in eq.(27).

¶ $m(h)$ is the inverse function of $\widehat{h}(\delta)$. That is:
A plot of h vs. $m(h)$
is identical to
A plot of $\widehat{h}(\delta)$ vs. δ.

¶ Knowledge of $m(h)$ is equivalent to knowledge of $\widehat{h}(\delta)$ because the robustness function is monotonic.

¶ We can express $m(h)$ by noting that:

$$\sigma(\varepsilon)-\widetilde{\sigma}(\varepsilon) = [\widetilde{\sigma}(\varepsilon)+s(\varepsilon)]-\widetilde{\sigma}(\varepsilon) = s(\varepsilon) = \beta^T\gamma(\varepsilon) \tag{28}$$

¶ Thus:

$$m(h) = \max_{\beta\in\mathcal{U}(h)}\beta^T\gamma(\varepsilon) \tag{29}$$

¶ Thus we must maximize $\beta^T\gamma(\varepsilon)$ subject to the constraint of the info-gap model at horizon of uncertainty h. This maximum occurs on the boundary of the ellipsoid. **(Why?)**

¶ We use **Lagrange optimization.** Define:

$$H = \beta^T\gamma(\varepsilon) + \lambda\left[h^2 - \left(\beta-\widetilde{\beta}\right)^T W\left(\beta-\widetilde{\beta}\right)\right] \tag{30}$$

We must maximize the 1st term on the right, while the 2nd term on the right equals 0. (**Why?**)

¶ Differentiate H with respect to β, equate to zero, solve for β (**will this give a maximum or minimum?**):

$$\frac{\partial H}{\partial\beta} = \gamma - 2\lambda W\left(\beta-\widetilde{\beta}\right) = 0 \tag{31}$$

$$\Longrightarrow \quad \beta-\widetilde{\beta} = \frac{1}{2\lambda}W^{-1}\gamma \tag{32}$$

$$\Longrightarrow \quad \beta = \widetilde{\beta} + \frac{1}{2\lambda}W^{-1}\gamma \tag{33}$$

¶ The constraint from the boundary of the info-gap model, using eq.(32), is:

$$h^2 = \left(\beta - \widetilde{\beta}\right)^T W \left(\beta - \widetilde{\beta}\right) \quad (34)$$

$$= \frac{1}{4\lambda^2}\gamma^T W^{-1} W W^{-1}\gamma \quad (35)$$

$$= \frac{1}{4\lambda^2}\gamma^T W^{-1}\gamma \quad (36)$$

$$\Longrightarrow \frac{1}{2\lambda} = \frac{\pm h}{\sqrt{\gamma^T W^{-1}\gamma}} \quad (37)$$

$$\Longrightarrow \beta = \widetilde{\beta} + \frac{\pm h}{\sqrt{\gamma^T W^{-1}\gamma}} W^{-1}\gamma \quad (38)$$

$$\Longrightarrow m(h) = \beta^T\gamma = \widetilde{\beta}^T\gamma + h\sqrt{\gamma^T W^{-1}\gamma} \leq \delta \quad (39)$$

- **Why** did I choose the '+' root in eq.(38)?
- **Why** does $\beta^T\gamma = \gamma^T\beta$?

¶ We derive the robustness by equating $m(h)$ to δ and solving for h:

$$m(h) = \delta \quad \Longrightarrow \quad \boxed{\widehat{h}(\delta) = \frac{\delta - \widetilde{\beta}^T\gamma}{\sqrt{\gamma^T W^{-1}\gamma}}} \quad (40)$$

or zero if this is negative. (**Why?**)

1.4 Robustness with Energy-Bound Functional Uncertainty

1.4.1 Formulation of the Robustness Function

¶ We now repeat the development in section 1.3, but with different:
• information about the uncertain function.
• info-gap model of uncertainty.

¶ Consider an uncertain function, $\sigma(\varepsilon)$, that may be stress as a function of strain, or any of the functional relations considered before, or any other functional relation.

¶ We represent $\sigma(\varepsilon)$, **the system model,** as in eq.(21):

$$\sigma(\varepsilon) = \widetilde{\sigma}(\varepsilon) + s(\varepsilon) \tag{41}$$

where:
$\widetilde{\sigma}(\varepsilon) =$ putative best estimate of the function.
$s(\varepsilon) =$ unknown error of the putative best estimate.

¶ **We know** several things about the uncertain function $s(\varepsilon)$:
• Its typical value is the known function $\widetilde{s}(\varepsilon)$, which may simply be zero.
• There can be **large transient deviations** between $s(\varepsilon)$ and $\widetilde{s}(\varepsilon)$.
• The deviation between $s(\varepsilon)$ and $\widetilde{s}(\varepsilon)$ is not persistent: it is asymptotically zero.

¶ We can express **large transient uncertainties** with the **energy-bound info-gap model:**

$$\mathcal{U}(h) = \left\{ s(\varepsilon) : \ \int [s(\varepsilon) - \widetilde{s}(\varepsilon)]^2 \, \mathrm{d}\varepsilon \le h^2 \right\}, \quad h \ge 0 \tag{42}$$

• **Why** does this info-gap model represent the potential for large transient uncertainty?

¶ Our **performance requirement** is that the weighted global value of $\sigma(\varepsilon)$ is less than a critical value:

$$\int \sigma(\varepsilon) w(\varepsilon) \, \mathrm{d}\varepsilon \le \delta \tag{43}$$

• This differs from the performance requirement in eq.(25), which focussed on a single value of ε.
• We denote the integral on the left hand side of eq.(43) as Δ, or $\widetilde{\Delta}$ if $\widetilde{\sigma}$ is used.

¶ In analogy to eq.(27), the **definition of the robustness function** is:

$$\widehat{h}(\delta) = \max \left\{ h : \ \left(\max_{s(\varepsilon) \in \mathcal{U}(h)} \int \sigma(\varepsilon) w(\varepsilon) \, \mathrm{d}\varepsilon \right) \le \delta \right\} \tag{44}$$

1.4.2 Derivation of the Robustness Function

¶ Let $m(h)$ denote the inner maximum in eq.(44). $m(h)$ is the inverse of $\widehat{h}(\delta)$.

¶ We use the **Cauchy-Schwarz inequality** to derive an expression for $m(h)$. For any functions $f(x)$ and $g(x)$:

$$\left(\int f(x)g(x)\,\mathrm{d}x\right)^2 \le \int f^2(x)\,\mathrm{d}x \int g^2(x)\,\mathrm{d}x \tag{45}$$

with equality if and only if $f(x)$ is proportional to $g(x)$, that is, $f(x) = cg(x)$ for non-zero c.

¶ We now derive an expression for $m(h)$. Note that:

$$\begin{aligned}
\int \sigma(\varepsilon)w(\varepsilon)\,\mathrm{d}\varepsilon &= \int [\widetilde{\sigma}(\varepsilon) + s(\varepsilon)]w(\varepsilon)\,\mathrm{d}\varepsilon && (46)\\
&= \int \widetilde{\sigma}(\varepsilon)w(\varepsilon)\,\mathrm{d}\varepsilon + \int s(\varepsilon)w(\varepsilon)\,\mathrm{d}\varepsilon && (47)\\
&= \int \widetilde{\sigma}(\varepsilon)w(\varepsilon)\,\mathrm{d}\varepsilon + \int [s(\varepsilon) - \widetilde{s}(\varepsilon)]w(\varepsilon)\,\mathrm{d}\varepsilon + \int \widetilde{s}(\varepsilon)w(\varepsilon)\,\mathrm{d}\varepsilon && (48)\\
&= \underbrace{\int [\widetilde{\sigma}(\varepsilon) + \widetilde{s}(\varepsilon)]w(\varepsilon)\,\mathrm{d}\varepsilon}_{\widetilde{\Delta}} + \int [s(\varepsilon) - \widetilde{s}(\varepsilon)]w(\varepsilon)\,\mathrm{d}\varepsilon && (49)\\
&\le \widetilde{\Delta} + \sqrt{\int [s(\varepsilon) - \widetilde{s}(\varepsilon)]^2\,\mathrm{d}\varepsilon}\sqrt{\int w^2(\varepsilon)\,\mathrm{d}\varepsilon} && (50)\\
&\le \widetilde{\Delta} + h\sqrt{\int w^2(\varepsilon)\,\mathrm{d}\varepsilon} && (51)
\end{aligned}$$

- The inequality in eq.(50) derives from Cauchy-Schwarz.
- The inequality in eq.(51) derives from the info-gap model, eq.(42).

¶ Thus $m(h)$ is:

$$m(h) = \widetilde{\Delta} + h\sqrt{\int w^2(\varepsilon)\,\mathrm{d}\varepsilon} \le \delta \tag{52}$$

- **Why** can $s(\varepsilon) - \widetilde{s}(\varepsilon)$ be chosen in eq.(50) so that the righthand side of eq.(51) is a least upper bound?

¶ We derive the **robustness function** by equating $m(h)$ to δ and solving for h:

$$m(h) = \widetilde{\Delta} + h\sqrt{\int w^2(\varepsilon)\,\mathrm{d}\varepsilon} = \delta \quad \Longrightarrow \quad \boxed{\widehat{h}(\delta) = \frac{\delta - \widetilde{\Delta}}{\sqrt{\int w^2(\varepsilon)\,\mathrm{d}\varepsilon}}} \tag{53}$$

or zero if this is negative. (**Why? What does this mean?**)

2 System Identification

¶ Optimal system identification: Adjusting a model to conform to data.

¶ **Main thesis:**
Optimal identification has no robustness to residual errors in the model.

¶ **Corollaries:**
• Sub-optimal models can be robust.
• Sub-optimal models can
○ be more robust than, and
○ reproduce data as well as,
the optimal model.

2.1 Model Uncertainty: Preliminary Example

¶ **Examples of model uncertainty:**
• Tracking elusive robotic swarm.
• Modeling material behavior in unmeasurable conditions (e.g. LANL).
• Modeling economic trends.

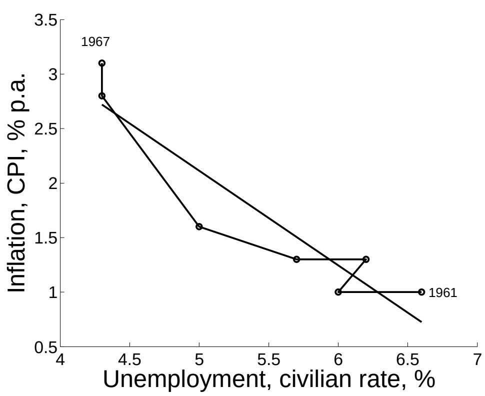


Figure 11: Inflation vs. unemployment in the US, 1961–1967.

Figure 12: Inflation vs. unemployment in the US, 1961–1993.

¶ From fig. 11, US unemployment vs. inflation for 1961–1967 looks linear:

$$\pi = aU + b \tag{54}$$

¶ From fig. 12 shows more complicated dynamics.
¶ Slopes in other periods:
○ '61–'67: -0.87 ○ '80–'83: -3.34 ○ '85–'93: -1.08 ○ '70–'78: ???

¶ Info-gaps:
○ Uncertain data and process.
○ Unknown functional relation.
¶ In section 1 we consider **uncertain data.** Now we consider **uncertain model structure.**

2.2 Optimal System Identification

¶ Notation:

y_i	= i th data set, $i = 1, \ldots, N$,
$f_i(q)$	= Model prediction of y_i.
q	= Parameters and properties of model.
$\mathcal{Y}$	= $\left\{ y_1, \ldots, y_N \right\}$.
$\mathcal{F}(q)$	= $\left\{ f_1(q), \ldots, f_N(q) \right\}$.
$R[\mathcal{Y}, \mathcal{F}(q)]$	= Performance of predictor, e.g. mean-square error:

$$R[\mathcal{Y}, f(q)] = \frac{1}{N} \sum_{i=1}^{N} \|f_i(q) - y_i\|^2 \tag{55}$$

¶ Optimal model, $q^\bullet$, minimizes performance-measure:

$$q^\bullet = \arg\min_q R[\mathcal{Y}, F(q)] \tag{56}$$

¶ We will show: fidelity of model to data as good as $R[\mathcal{Y}, f(q^\bullet)]$ is

- obtainable but not reliable if there are info-gaps.
- not robust to info-gaps in model.

2.3 Uncertainty

¶ Model structure $f_i(q)$ is wrong. Relevant factors are missing:
- Non-linearities.
- Time dependence.
- Dimensionality.
- Etc.

¶ Complete model:

$$\phi_i = f_i(q) + u_i \tag{57}$$

$f_i(q)$ = Best known model structure.
ϕ_i = Correct model structure.
u_i = Unknown info-gap.

¶ Info-gap model of uncertainty: Unbounded family of nested sets (of models):

$$f_i(q) \in \mathcal{U}(h, f_i(q)), \quad h \geq 0 \tag{58}$$

$$h < h^{\bullet} \implies \mathcal{U}(h, f_i(q)) \subset \mathcal{U}\left(h^{\bullet}, f_i(q)\right) \tag{59}$$

2.4 Robustness

¶ Fidelity of model to data:
$R[\mathcal{Y}, \mathcal{F}(q)]$ = Fidelity of model $f_i(q)$ to data.
$R[\mathcal{Y}, \mathcal{F}_u(q)]$ = Fidelity of model $f_i(q) + u_i$ to data.
r_{c} = Acceptable fidelity of model to data.

¶ Robustness of model $f_i(q)$:

- How wrong can $f_i(q)$ be without exceeding acceptable fidelity?
- Epistemic, not ontological question.
- Max horizon of uncertainty, h, which does not jeopardize fidelity:

$$\widehat{h}(q, r_{\mathrm{c}}) = \max \left\{ h : \max_{\substack{\phi_i \in \mathcal{U}(h, f_i(q)) \\ i=1,\ldots,N}} R[\mathcal{Y}, \mathcal{F}_u(q)] \le r_{\mathrm{c}} \right\} \tag{60}$$

2.5 Performance and Robustness

¶ $R[\mathcal{Y}, \mathcal{F}(q)]$ = Fidelity of model, $f_i(q)$, to data.

¶ $\widehat{h}(q, r_{\mathrm{c}})$ = Robustness of model, $f_i(q)$, with fidelity-aspiration r_{c}.
¶ Theorem:

$$r_{\mathrm{c}} = R[\mathcal{Y}, \mathcal{F}(q)] \quad \text{implies} \quad \widehat{h}(q, r_{\mathrm{c}}) = 0 \tag{61}$$

Meaning:
No model can be relied upon to perform “as advertised”.

¶ This holds also for optimal model, $q^\bullet$:

$$R[\mathcal{Y}, f(q^\bullet)] = \min_q R[\mathcal{Y}, f(q)] \tag{62}$$

$$R_{\mathrm{C}}^\bullet = R[\mathcal{Y}, \mathcal{F}(q^\bullet)] \quad \text{implies} \quad \widehat{h}(q^\bullet, R_{\mathrm{C}}^\bullet) = 0 \tag{63}$$

¶ Implication:
Sub-optimal models can be more robust than optimal model at same fidelity.

2.6 Example

¶ 1-dimensional system:

y_i = Scalar measurements.
$f_i(q)$ = qi. Nominal linear model
$R[\mathcal{Y}, \mathcal{F}(q)]$ = Mean-squared error:

$$R[\mathcal{Y}, f(q)] = \frac{1}{N} \sum_{i=1}^{N} (qi - y_i)^2 \tag{64}$$

¶ $q^\bullet$ = Least-squares optimal model:

$$q^\bullet = \arg\min_q R[\mathcal{Y}, f(q)] = \frac{\eta_1}{\eta_0} \tag{65}$$

$$\eta_1 = \frac{1}{N} \sum_{i=1}^{N} i\, y_i, \quad \eta_0 = \frac{1}{N} \sum_{i=1}^{N} i^2 \tag{66}$$

¶ Model error: Uncertain quadratic term.

$$\phi_i = qi \ + \ ui^2 \tag{67}$$

¶ Info-gap model for quadratic uncertainty:

$$\mathcal{U}(h, qi) = \left\{ \phi_i = qi \ + \ ui^2 : \ |u| \le h \right\}, \quad h \ge 0 \tag{68}$$

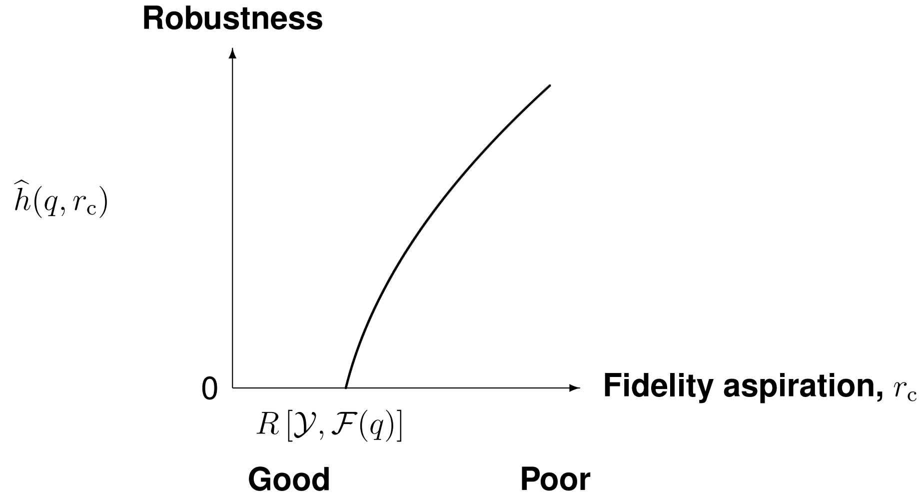
¶ **Robustness:**
Max horizon of uncertainty, h, with acceptable fidelity to data.

$$\widehat{h}(q, r_{\rm c}) = \max \left\{ h : \max_{|u| \, \le \, h} R \left[\mathcal{Y}, \mathcal{F}_u(q) \right] \le r_{\rm c} \right\} \tag{69}$$

$$\widehat{h}(q, r_{\rm c}) = \begin{cases} 0 \, , & r_{\rm c} \le \xi_2 \\ \frac{|\xi_1|}{\xi_0} \left(-1 + \sqrt{1 + \frac{r_{\rm c} - \xi_2}{\xi_1^2}} \right), & \xi_2 < r_{\rm c} \end{cases} \tag{70}$$

$$\begin{aligned} \xi_2 &= \frac{1}{N} \sum_{i=1}^{N} (qi \; - \; y_i)^2 && (71) \\ &= R \left[\mathcal{Y}, \mathcal{F}(q) \right] && (72) \\ \xi_1 &= \frac{1}{N} \sum_{i=1}^{N} i^2 \; (qi \; - \; y_i) && (73) \\ \xi_0 &= \frac{1}{N} \sum_{i=1}^{N} i^4 && (74) \end{aligned}$$

¶ **Trade-off: robustness vs. fidelity.**

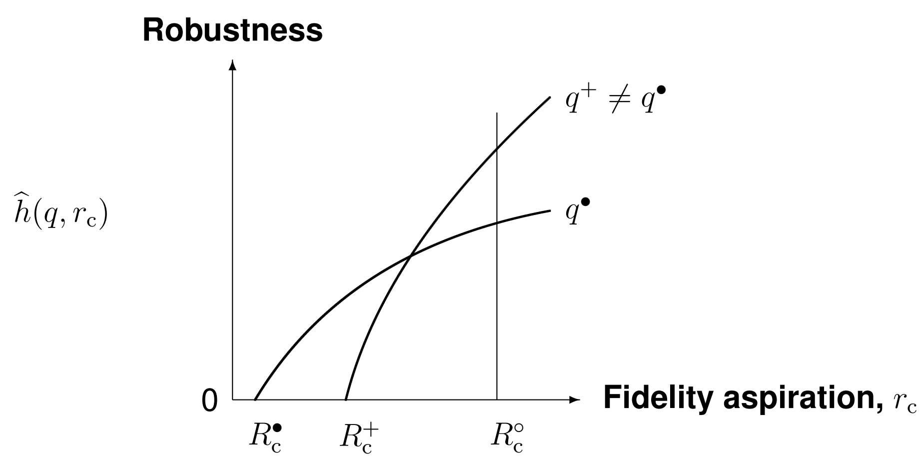


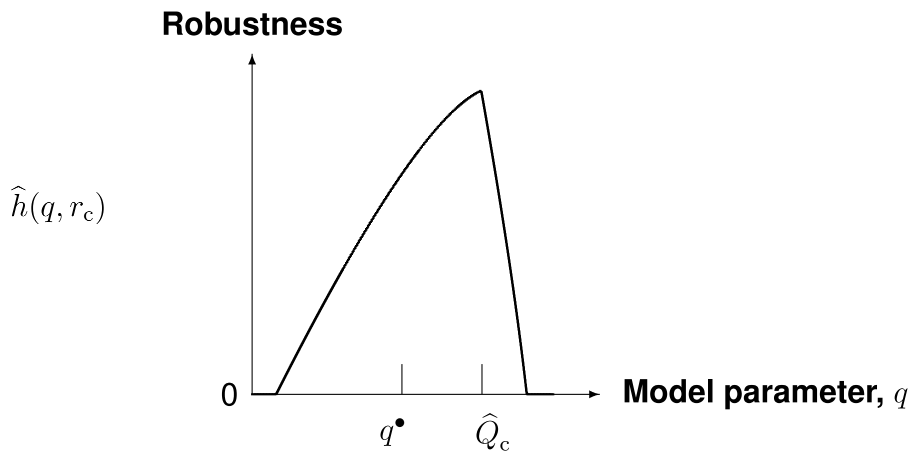
¶ **No robustness for** aspiration at nominal performance:

$$\widehat{h}(q, r_c) = 0 \quad \textbf{if} \quad r_c = R[\mathcal{Y}, \mathcal{F}(q)] \tag{75}$$

¶ **Preference for sub-optimal model:**

$q^\bullet$ = L.S.-optimal model.
q^+ = L.S.-sub-optimal model.
R_C° = Acceptable fidelity.
q^+ preferred to $q^\bullet$ at R_C°.





¶ **Robust-satisficing model:**

$q^{\bullet}$ = L.S.-optimal model. $R^{\bullet}$ = L.S. optimal error.
$\widehat{q}_{\rm c}$ =Robust-satisficing model. Maximizes $\widehat{h}$(q , $r_{\rm c}$).
$R_{\rm C}$ only slightly $> R^{\bullet}$. $\widehat{h}(\widehat{q}_{\rm c}, r_{\rm c}) \gg \widehat{h}(q^{\bullet}, r_{\rm c})$.
$\widehat{q}_{\rm c}$ preferred to $q^{\bullet}$.

¶ Conclusions:

- Any model, $f_i(q)$,
  - has no immunity to unknown quadratic term:
    $$\widehat{h}(q, r_{\rm c}) = 0 \quad \text{if} \quad r_{\rm c} = R\,[\mathcal{Y}, \mathcal{F}(q)].$$
  - is reliable only at less-than-nominal fidelity.

- Also holds for least-square optimal model, $q^{\bullet}$.

- Robustness curves can cross:
  Sub-optimal model q^{+}
  more robust than optimal model $q^{\bullet}$
  at same fidelity to data.

- Info-gap strategy:
  - Satisfice fidelity to data.
  - Optimize robustness to model-deficiency.

2.7 An Interpretation: Focus of Uncertainty

¶ **Least-squares estimation** focusses on managing error in data, y_i:

$$\textbf{Minimize:} \quad \sum_{i=1}^{N} (f_i(q) - y_i)^2 \tag{76}$$

¶ **Info-gap estimation** focusses on managing

- error in data, y_i:

  Satisfice: $\sum_{i=1}^{N} (f_i(q) - y_i)^2$

- error in model, $f_i(q)$:

  Maximize: $\widehat{h}(q, r_{\rm c})$.

2.8 Robustness and Opportuneness

¶ **Robustness of model** $f_i(q)$:

how wrong can $f_i(q)$ be without exceeding acceptable fidelity?

$$\widehat{h}(q, r_{\rm c}) = \max \left\{ h : \max_{\substack{\phi_i \in \mathcal{U}(h, f_i(q)) \\ i=1,\ldots,N}} R\left[\mathcal{Y}, \mathcal{F}_u(q)\right] \le r_{\rm c} \right\} \qquad (77)$$

¶ **Opportuneness of model** $f_i(q)$:

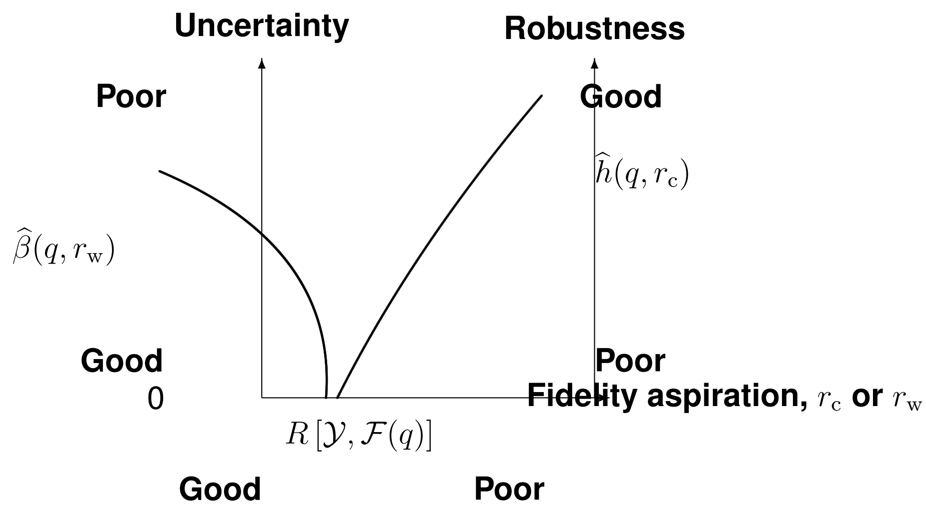
how wrong must $f_i(q)$ be to enable windfall fidelity?

$$r_{\rm w} \ll r_{\rm c}$$

$$\widehat{\beta}(q, r_{\rm w}) = \min \left\{ h : \min_{\substack{\phi_i \in \mathcal{U}(h, f_i(q)) \\ i=1,\ldots,N}} R\left[\mathcal{Y}, \mathcal{F}_u(q)\right] \le r_{\rm w} \right\} \qquad (78)$$

¶ Preferences:

- Robustness:
  - ◦ Immunity to failure.
  - ◦ Satisficing at critical fidelity.
  - ◦ Bigger is better
- Opportuneness:
  - ◦ Immunity to windfall.
  - ◦ Windfalling at wildest-dream fidelity.
  - ◦ Big is bad.



¶ **Trade-offs:**

- Robustness vs. critical fidelity.
- Opportuneness vs. windfall fidelity.

¶ Sympathetic immunities:
change in model, q , which improves $\widehat{h}$
also improves $\widehat{\beta}$:

$$\frac{\partial\widehat{h}}{\partial q}\frac{\partial\widehat{\beta}}{\partial q} < 0 \tag{79}$$

¶ Antagonistic immunities:
change in model, q , which improves $\widehat{h}$
also degrades $\widehat{\beta}$:

$$\frac{\partial\widehat{h}}{\partial q}\frac{\partial\widehat{\beta}}{\partial q} > 0 \tag{80}$$

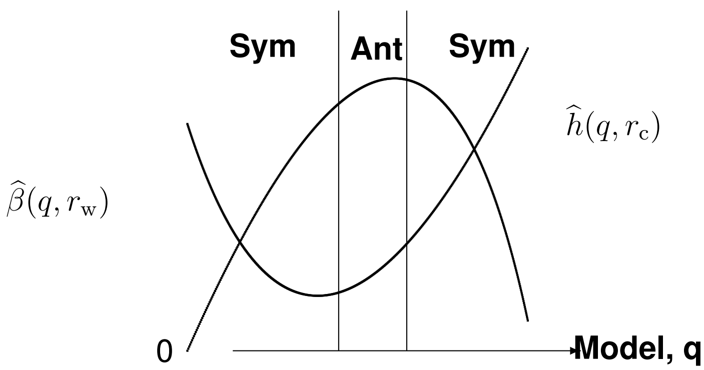


Figure 13: **Schematic immunity curves**

$$\left(\frac{\widehat{h}}{|u_M|}+1\right)^2 - \frac{\xi_0 r_c}{\xi_1} = \left(\frac{\widehat{\beta}}{|u_M|}-1\right)^2 - \frac{\xi_0 r_w}{\xi_1} \tag{81}$$

2.9 Forecasting and looseness of model prediction

¶ **Source:** Yakov Ben-Haim and Francois Hemez, 2011, Robustness, Fidelity and Prediction-Looseness of Models, *Proceedings of the Royal Society, A,* to appear.

¶ **The issue of prediction looseness:**

- At high robustness,
  Many models have same fidelity (because they are all sub-optimal).
- Do their predictions agree?

¶ **Unknown complete model:**

$$\phi_i = f_i(q) + u_i \tag{82}$$

¶ The info-gap uncertainty model is:

$$\mathcal{U}[h, f_i(q)], \quad h \geq 0.$$

¶ For design q define:

$h^\star$ $= \widehat{h}(q, r_{\rm c})$
= Robustness of q at $r_{\rm c}$.
$\Lambda(q)$ = $\mathcal{U}[h^\star, f_i(q)]$
= set of all models, ϕ_i, which satisfice the prediction error at $r_{\rm c}$.
= Predictions of fidelity-equivalent models.
= **Prediction-looseness** of model q.

¶ **Fidelity–robustness trade-off:**

$$r_{\rm c} < R_{\rm C}^{\bullet} \quad \Longrightarrow \quad \widehat{h}(q, r_{\rm c}) \leq \widehat{h}(q, R_{\rm C}^{\bullet}) \tag{83}$$

Robustness decreases as fidelity improves.

¶ **Robustness–prediction-looseness trade-off:**

$$\widehat{h}(q, r_{\rm c}) < \widehat{h}(q^{\bullet}, r_{\rm c}) \quad \Longrightarrow \quad \Lambda(q) \subseteq \Lambda(q^{\bullet}) + \mu \tag{84}$$

Robustness decreases as looseness improves.

¶ **The dilemma:**
- Fidelity to data necessary for trueness of model.
- Robustness to model uncertainty verifies fidelity.
- Looseness of model prediction results from
  fidelity-robustness to model-uncertainty.

¶ Dilemma due to **conflict of two uncertainties:**
- Measurement error (spread of data).
  Causes need for fidelity.
- Model error (epistemic limitation).
  Causes need for robustness.

¶ **Hume and the problem of induction:**
- The past does not bind the future.
- Experience cannot validate scientific induction.

¶ **Robustness-fidelity-looseness trade-offs:**
Measurement error and limited understanding
impose prediction looseness.

¶ **Epistemological warrant:**
- Basis for theory (model) selection.
- Obtained by:
  - High fidelity to data.
  - High robustness to model error.

¶ Question: **Is warrant warranted?**
Warrant = Hi fidelity and high robustness
= High prediction looseness.
Answer: Doesn't look like it.

3 Tychonov Up-Dating of a Linear System with Model Uncertainty

3.1 Formulation of the Up-Dating Problem

¶ **Measurements:**
$f \in \Re^J$ is the exact force vector applied to a system.
$y^{(m)} \in \Re^N$ is the noisy response vector, for experiment $m = 1, \ldots, M$.

¶ **Model we will up-date:** choose the flexibility matrix V in:

$$y = Vf \tag{85}$$

¶ **Ill-conditioning:**
• The mean squared error is:

$$S = \frac{1}{M}\sum_{m=1}^{M} \|y^{(m)} - y\|^2 \tag{86}$$

$$= \frac{1}{M}\sum_{m=1}^{M} \|y^{(m)} - Vf\|^2 \tag{87}$$

$$= \frac{1}{M}\sum_{m=1}^{M} \left(y^{(m)} - Vf\right)^T \left(y^{(m)} - Vf\right) \tag{88}$$

• The least squares estimate is the choice of V that satisfies:

$$\frac{\partial S}{\partial V} = 0 \tag{89}$$

◦ This is very sensitive to noise in the observations, $y^{(m)}$ and f (we assume f is known).
◦ One approach to handle this noise-sensitivity is called Tychonov regularization.

¶ **Tychonov-regularized least squared error** is:

$$S = \lambda\|\widetilde{y} - y||^2 + \frac{1}{M}\sum_{m=1}^{M} \|y^{(m)} - y\|^2 \tag{90}$$

where:
• $\widetilde{y}$ is a prior estimate of the response.
• λ is a weighting term. Ordinary LS: $\lambda = 0$.
• We are using the Euclidean norm: $\|x\|^2 = x^T x$.

¶ **Two uncertainties:**
• Statistical: noisy data.
• Info-gap: uncertain model structure. Specifically, inhomogeneous input/output relation:

$$y = Vf + u \tag{91}$$

[1]This section based on:
Yakov Ben-Haim and Scott Cogan, Up-Dating a Linear System with Model Uncertainty: An Info-Gap Approach, Intl. Conf. on Uncertainty in Structural Dynamics, University of Sheffield, UK. 15–17.6.2009.

The data don't reflect this info-gap. E.g. Lab vs real-life, change due to wear, ignorance, etc.

¶ **Actual mean-squared error.** Substituting eq.(91) into eq.(90):

$$\begin{aligned} S(V,u) &= \underbrace{(1+\lambda)f^TV^TVf - 2\left(\lambda\widetilde{y}+\overline{y}\right)^T Vf + \lambda\|\widetilde{y}\|^2 + \overline{\|y\|^2}}_{S_\mathrm{o}} \\ &\quad \underbrace{+\,(1+\lambda)u^Tu - \underbrace{2\left(\lambda\widetilde{y}+\overline{y}-(1+\lambda)Vf\right)^T u}_{2z^Tu}}_{S_\mathrm{u}} \qquad (92) \\ &= S_\mathrm{o} + (1+\lambda)u^Tu - 2z^Tu \qquad (93) \end{aligned}$$

where:

$$\overline{y} = \frac{1}{M}\sum_{m=1}^{M} y^{(m)} \qquad (94)$$

$$\overline{\|y\|^2} = \frac{1}{M}\sum_{m=1}^{M} \|y^{(m)}\|^2 \qquad (95)$$

• S_o is the ordinary Tychonov-regularized least-squares error function for the linear model, $y = Vf$.

• S_u contains the uncertain inhomogeneous terms in the model in eq.(91). S_u also contains the measurements, f and $y^{(1)}, \ldots, y^{(M)}$, in the vector z and in S_o.

¶ **Goal.**

• We wish to choose V

• We cannot actually minimize $S(V,u)$ since u is unknown.

• The approach: choose V to make $S(V,u)$
adequately small for a **maximal range** of possible realizations of u.

• **Key concept: robust satisficing.**

• **What** are we optimizing? **What** are we satisficing?

3.2 Robustness to Uncertainty

¶ **System model:** $S(V,u)$ in eq.(93).

¶ **Uncertainty model:** spherical info-gap model for uncertain vector u in eq.(91):

$$\mathcal{U}(h) = \left\{u:\ u^Tu \le h^2\right\}, \quad h \ge 0 \qquad (96)$$

¶ **Performance requirement:** regularized squared error must not exceed S_c:

$$S(V,u) \le S_\mathrm{c} \qquad (97)$$

¶ **Robustness function.**

$$\widehat{h}(V,S_\mathrm{c}) = \max\left\{h:\ \left(\max_{u\in\mathcal{U}(h)} S(V,u)\right) \le S_\mathrm{c}\right\} \qquad (98)$$

¶ We will soon show that:

$$\widehat{h}(V, S_c) = \frac{1}{1+\lambda}\left(-\sqrt{z^T z} + \sqrt{z^T z + (1+\lambda)(S_c - S_o)}\right) \tag{99}$$

or zero if $S_c \le S_o$.

• The dependence of the robustness on the model matrix, V, and on the observations f and $y^{(m)}$, arises through S_o and z, defined in eq.(92).

• Note **zeroing** and **trade off.**

¶ **Derivation of eq.(99):**

• We will use Lagrange optimization to evaluate $m(h)$, the inner maximum in eq.(98).

• We must maximize S in eq.(93) on p.31:

$$S = S_o + (1+\lambda)u^T u - 2z^T u \tag{100}$$

subject to the constraint that $u \in \mathcal{U}(h)$, eq.(96), p.31.

• By completing the square and comparing with eq.(100) we see that S is a spheroid.

• We must find v and Δ so that:

$$S = \overbrace{(1+\lambda)(u-v)^T(u-v)}^{S'} + \Delta \tag{101}$$

$$= (1+\lambda)u^T u - 2(1+\lambda)v^T u + (1+\lambda)v^T v + \Delta \tag{102}$$

$$\Longrightarrow \quad (1+\lambda)v = z \quad \Longrightarrow \quad v = \frac{1}{1+\lambda}z \tag{103}$$

$$\Longrightarrow \quad (1+\lambda)v^T v = \frac{1}{1+\lambda}z^T z \quad \Longrightarrow \quad S_o = \Delta + \frac{1}{1+\lambda}z^T z \tag{104}$$

$$\Longrightarrow \quad \Delta = S_o - \frac{1}{1+\lambda}z^T z \tag{105}$$

• To evaluate the robustness we must maximize S' subject to $u^T u \le h^2$.

◦ $S' = x^2$ is the set of u's that form a spheroid surface centered at v and of radius x.

◦ $u^T u \le h^2$ is the set of uncertain u's that form a solid sphere centered at the origin.

◦ S' is maximized, at fixed h, when the spheroid surface contains the solid sphere, and any further expansion of S' would no longer intersect the solid sphere: fig. 14.

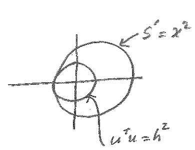


Figure 14: Intersection of spheroid surface, $S' = x^2$, with solid sphere, $u^T u \le h^2$.

◦ Thus S' is maximized by a u on the surface of the spheroid $u^T u = h^2$.

◦ Thus we can maximize subject to the equality constraint, $u^T u = h^2$.

• Define the objective function with Lagrange multiplier α, from S in eq.(93) on p.31:

$$H = \underbrace{S_o + (1+\lambda)u^T u - 2z^T u}_{S} + \alpha(h^2 - u^T u) \tag{106}$$

• The condition for an extremum is:

$$0 = \frac{\partial H}{\partial u} = 2(1+\lambda)u - 2\alpha u - 2z \tag{107}$$

$$\Longrightarrow \quad (1+\lambda-\alpha)u = z \tag{108}$$

$$\Longrightarrow \quad u = \frac{1}{1+\lambda-\alpha} z \tag{109}$$

• From the constraint, $u^T u = h^2$:

$$h^2 = \frac{1}{(1+\lambda-\alpha)^2} z^T z \quad \Longrightarrow \quad \frac{1}{1+\lambda-\alpha} = \frac{\pm h}{\sqrt{z^T z}} \quad \Longrightarrow \quad u = \frac{\pm h}{\sqrt{z^T z}} z \tag{110}$$

• Hence, from eq.(100), the inner maximum is:

$$m(h) = S_{\rm o} + (1+\lambda)h^2 \mp 2h\sqrt{z^T z} \tag{111}$$

Choose the '+' for a maximum.

• Equate $m(h)$ to $S_{\rm c}$ and solve for h to find the robustness:

$$m(h) = S_{\rm c} \quad \Longrightarrow \quad (1+\lambda)h^2 + 2h\sqrt{z^T z} + \underbrace{S_{\rm o} - S_{\rm c}}_{<0} = 0 \tag{112}$$

$$\Longrightarrow \quad h^2 + \frac{2\sqrt{z^T z}}{1+\lambda} h + \frac{S_{\rm o} - S_{\rm c}}{1+\lambda} = 0 \tag{113}$$

The coefficients of h change sign once so, by the Descartes rule,[2] there is 1 positive root.

• The positive root of eq.(113) is eq.(99), p.32.

3.3 Robustness of the Tychonov Regularized Model

¶ **Preview.** In this section we:

• Derive an explicit expression for the robustness of an up-dated model that minimizes the Tychonov-regularized mean squared error, $S_{\rm o}$ in eq.(92).

• Theorem 1 asserts that Tychonov-optimal matrices are more robust to uncertainty than all other matrices, at fixed Tychonov weight. (**Why** is this **Wow!**? More later.)

• Theorem 2 asserts that the robustness of Tychonov optimal matrices increases as the Tychonov weight decreases. (**Good news** or **bad news**? More later.)

• Proofs appear in appendix 3.5.

¶ **Tychonov-regularized mean squared error,** $S_{\rm o}$ in eq.(92), p.31, can be written:

$$S_{\rm o}(V) = (1+\lambda)\left[(Vf-a)^T(Vf-a)\right] + b \tag{114}$$

where a and b depend on the measurements:

$$a = \frac{1}{1+\lambda}(\lambda\widetilde{y} + \overline{y}) \tag{115}$$

$$b = \lambda\|\widetilde{y}\|^2 + \overline{\|y\|^2} - (1+\lambda)a^T a \tag{116}$$

[2]Pearson, Carl E., ed., *Handbook of Applied Mathematics.* 1st ed., p.11

¶ **Minimization of $S_{\rm o}(V)$:**

• If $f \neq 0$ then the matrix V can always be chosen to precisely satisfy $Vf = a$, which minimizes $S_{\rm o}(V)$ in eq.(114).

• Let $V_{\rm T}$ denote any such choice of V, which we will refer to as a *Tychonov optimal matrix*.

• It then results that z, defined in eq.(92), is identically zero.

• Furthermore one finds that $S_{\rm o}(V_{\rm T}) = b$.

• One now finds the robustness in eq.(99), for any Tychonov optimal matrix $V_{\rm T}$, to be:

$$\widehat{h}(V_{\rm T}, S_{\rm c}) = \sqrt{\frac{S_{\rm c} - b}{1+\lambda}} \tag{117}$$

or zero if $S_{\rm c} \le b$.

Theorem 1 *A Tychonov optimal matrix, $V_{\rm T}$, is strictly more robust than any other matrix V, at fixed Tychonov weight λ:*

$$\widehat{h}(V_{\rm T}, S_{\rm c}) > \widehat{h}(V, S_{\rm c}) \tag{118}$$

for all values of $S_{\rm c} > b$.

¶ Note relation to result by Zacksenhouse *et al:*
Zacksenhouse *et al.*[3] [proposition 2] derive a similar result though they consider info-gap uncertainty in the data, rather than uncertainty in the model structure as we have done.

Theorem 2 *Robustness of a Tychonov optimal matrix decreases as the Tychonov weight increases.*
Given two Tychonov weights, $\lambda < \lambda'$, with corresponding Tychonov optimal matrices $V_{\rm T}$ and $V'_{\rm T}$, respectively. Then:

$$\widehat{h}(V_{\rm T}, S_{\rm c}) > \widehat{h}(V'_{\rm T}, S_{\rm c}) \tag{119}$$

for all values of $S_{\rm c} > b(\lambda)$.

The proof of this theorem depends on the following lemma. First define the variance of the measured responses as:

$$\overline{\|y - \overline{y}\|^2} = \overline{\|y\|^2} - \|\overline{y}\|^2 \tag{120}$$

where the two terms on the right are defined in eqs.(94) and (95).

Lemma 1 *The coefficient b in eq.(95) can be expressed:*

$$b = \frac{\lambda}{1+\lambda} \|\widetilde{y} - \overline{y}\|^2 + \overline{\|y - \overline{y}\|^2} \tag{121}$$

¶ Note from lemma 1 that:

$$\widetilde{y} = \overline{y} \quad \text{implies} \quad b = \overline{\|y - \overline{y}\|^2} \tag{122}$$

Thus, if the Tychonov estimate, $\widetilde{y}$, equals the measured average, $\overline{y}$, then:

• b is independent of λ.

• $\widehat{h}(V_{\rm T}, S_{\rm c})$ decreases with increasing λ, but does not shift to the right.

[3]Zacksenhouse, M., S.Nemets, M.A.Lebedev and M.A.L.Nicolelis, 2009, Robust-satisficing linear regression: Performance/robustness trade-off and consistency criterion, *Mechanical Systems and Signal Processing,* 23: 1954–1964.

¶ Theorems 1 and 2 are illustrated in figs. 15 and 16. The data are in the footnotes[4] and[5].

• **Tychonov optimal matrices are more robust than other matrices** (evaluated at the same Tychonov weight): fig. 15.

BUT

• **Increasing** the Tychonov weight **reduces** the robustness of the Tychonov optimal matrix: fig. 16.

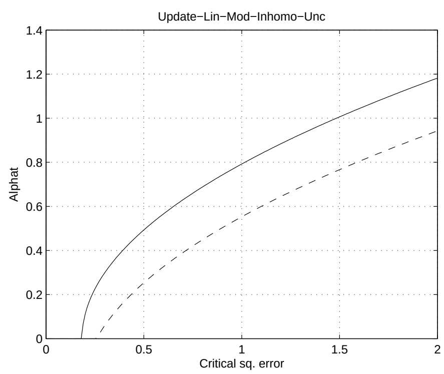


Figure 15: Robustness curves illustrating theorem 1. Tychonov-optimal (solid) and a different V matrix (dash). Tychonov weight: $\lambda = 0.3$

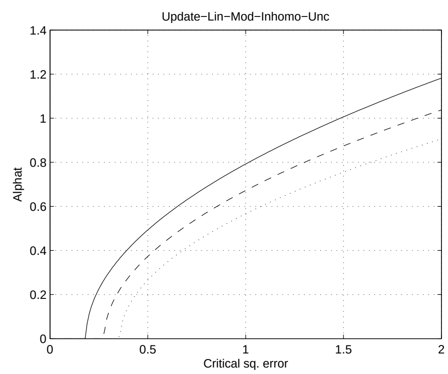


Figure 16: Robustness curves illustrating theorem 2. Tychonov optimal matrices with different weights: $\lambda = 0.3$ (solid), 0.6 (dash) and 1.0 (dot).

¶ **Implications of the theorems:**

• Theorem 1: Tychonov better (more robust) than non-Tychonov.

• Theorem 2: Less Tychonov better (more robust) than more Tychonov.

3.4 Recap of Tychonov Updating and Robustness

Let us recap the main ideas of this section.

First idea:

(1) Linear system estimation with noisy measurements and
(2) Matrix ill-conditioning
are resolved by
(3) Tychonov regularization.

Second idea:

(1) Model-structure uncertainty and
(2) Info-gap robustness
lead to:

[4]The data for these figures are:

$$\widetilde{y}^T = (2.3 \;\; 1.2), \quad f^T = (1 \;\; 0.7 \;\; 0.3)$$
$$Y = \begin{pmatrix} 3.0 & 3.2 & 2.8 & 3.1 \\ 1.5 & 1.4 & 1.6 & 1.7 \end{pmatrix}$$

[5]The non-Tychonov matrix is $V = \begin{pmatrix} 1.2 & 1.6 & 2.5 \\ 0.5 & 0.9 & 1.5 \end{pmatrix}$.

(3a) Tychonov optimal matrix is most robust.
(3b) Robustness of Tychonov optimal matrix goes **up** as Tychonov weight goes **down**.

3.5 Proofs

Proof of theorem 1. Since λ is non-negative, we see from eq. (114) that:

$$b \le S_{\rm o}(V) \tag{123}$$

with strict inequality unless V is itself a Tychonov optimal matrix.
Hence, since V is *not* a Tychonov optimal matrix:

$$S_{\rm c} - S_{\rm o} < S_{\rm c} - b \tag{124}$$

for all values of $S_{\rm c}$. Hence:

$$(1+\lambda)(S_{\rm c} - S_{\rm o}) < (1+\lambda)(S_{\rm c} - b) \tag{125}$$

Thus:

$$z^T z + (1+\lambda)(S_{\rm c} - S_{\rm o}) < z^T z + (1+\lambda)(S_{\rm c} - b) \tag{126}$$

Hence:

$$z^T z + (1+\lambda)(S_{\rm c} - S_{\rm o}) \;<\; z^T z + \sqrt{z^T z}\sqrt{(1+\lambda)(S_{\rm c} - b)} + (1+\lambda)(S_{\rm c} - b) \tag{127}$$

$$= \left(\sqrt{z^T z} + \sqrt{(1+\lambda)(S_{\rm c} - b)} \right)^2 \tag{128}$$

Thus:

$$\sqrt{z^T z + (1+\lambda)(S_{\rm c} - S_{\rm o})} < \sqrt{z^T z} + \sqrt{(1+\lambda)(S_{\rm c} - b)} \tag{129}$$

Hence

$$\frac{1}{1+\lambda}\left(-\sqrt{z^T z} + \sqrt{z^T z + (1+\lambda)(S_{\rm c} - S_{\rm o})} \right) < \sqrt{\frac{S_{\rm c} - b}{1+\lambda}} \tag{130}$$

which, by referring to eqs.(99) and (117) and recalling that $S_{\rm c} > b$, proves the result. ∎
Proof of lemma 1. Combining eqs.(115) and (116) we can write:

$$b \;=\; \lambda\|\widetilde{y}\|^2 + \overline{\|y\|^2} - \frac{1}{1+\lambda}\left(\lambda^2\|\widetilde{y}\|^2 + 2\lambda\widetilde{y}^T\overline{y} + \|\overline{y}\|^2 \right) \tag{131}$$

$$= \frac{\lambda}{1+\lambda}\|\widetilde{y}\|^2 - \frac{2\lambda}{1+\lambda}\widetilde{y}^T\overline{y} + \overline{\|y\|^2} - \frac{1}{1+\lambda}\|\overline{y}\|^2 \tag{132}$$

Completing the square in the first two terms in eq.(132):

$$b \;=\; \frac{\lambda}{1+\lambda}\left(\|\widetilde{y}\|^2 - 2\widetilde{y}^T\overline{y} + \|\overline{y}\|^2 \right) - \frac{\lambda}{1+\lambda}\|\overline{y}\|^2 + \overline{\|y\|^2} - \frac{1}{1+\lambda}\|\overline{y}\|^2 \tag{133}$$

$$= \frac{\lambda}{1+\lambda}\|\widetilde{y} - \overline{y}\|^2 + \overline{\|y\|^2} - \|\overline{y}\|^2 \tag{134}$$

which, with the definition in eq.(120), completes the proof. ∎
Proof of theorem 2. Eqs.(117) and (121) enable explicit derivation of the partial derivative of $\widehat{h}(V_{\rm T}, S_{\rm c})$ with respective to λ, which is found to be strictly negative for all values of $S_{\rm c}$ for which the robustness is positive ($S_{\rm c} > b$):

$$\frac{\partial \widehat{h}(V_{\rm T}, S_{\rm c})}{\partial \lambda} = -\frac{\|\widetilde{y} - \overline{y}\|^2 + (1+\lambda)(S_{\rm c} - b)}{2(1+\lambda)^3}\sqrt{\frac{1+\lambda}{S_{\rm c} - b}} \tag{135}$$

∎

4 Estimating an Uncertain Probability Density

¶ The problem:

• Estimate parameters of a probability density function (pdf) based on observations.

• Common approach: select parameter values to maximize the likelihood function for the class of pdfs.

• In this section: simple example of a situation where the **form** of the pdf is uncertain, not only **parameters**.

¶ Notation:

• $x =$ random variable.

• $X = (x_1, \ldots, x_N) =$ random sample.

• $\widetilde{p}(x|\lambda) =$ be a pdf for x with parameters λ.

¶ Likelihood function:

$$L(\widetilde{p}) = \prod_{i=1}^{N} \widetilde{p}(x_i|\lambda) \tag{136}$$

¶ Maximum likelihood estimate (MLE):

$$\lambda^\star = \arg\max_{\lambda} L(X, \widetilde{p}) \tag{137}$$

¶ Examples of MLE.

• **Exponential distribution:** The pdf is:

$$\widetilde{p}(x|\lambda) = \lambda \mathrm{e}^{-\lambda x}, \quad x \geq 0 \tag{138}$$

The likelihood function, from eq.(136), is:

$$L(\widetilde{p}) = \prod_{i=1}^{N} \widetilde{p}(x_i|\lambda) = \lambda^N \exp\left(-\lambda \sum_{i=1}^{N} x_i\right) \tag{139}$$

Thus:

$$\frac{\partial L}{\partial \lambda} = \left(N\lambda^{N-1} - \lambda^N \sum_{i=1}^{N} x_i\right) \exp\left(-\lambda \sum_{i=1}^{N} x_i\right) \tag{140}$$

Equating to zero and solving for λ yields the MLE:

$$0 = \frac{\partial L}{\partial \lambda} \quad \Longrightarrow \quad 0 = N\lambda^{N-1} - \lambda^N \sum_{i=1}^{N} x_i \quad \Longrightarrow \quad \boxed{\frac{1}{\lambda_{\mathrm{MLE}}} = \frac{1}{N} \sum_{i=1}^{N} x_i} \tag{141}$$

Note that:

$$\mathrm{E}(x) = \frac{1}{\lambda_{\mathrm{MLE}}} \tag{142}$$

• **Normal distribution: MLE of the mean.** The pdf is:

$$\widetilde{p}(x|\lambda) = \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-(x-\mu)^2/2\sigma^2} \tag{143}$$

The likelihood function, from eq.(136), is:

$$L(\widetilde{p}) = \prod_{i=1}^{N} \widetilde{p}(x_i|\lambda) = \frac{1}{(2\pi)^{N/2}\sigma^N} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2\right) \quad (144)$$

Note that:

$$\mu_{\rm MLE} = \arg\max_{\mu} L = \arg\min_{\mu} \sum_{i=1}^{N}(x_i-\mu)^2 = \text{Least Squares Estimate} \quad (145)$$

Thus MLE and LSE agree. Define the squared error:

$$S = \sum_{i=1}^{N}(x_i-\mu)^2 \quad (146)$$

Thus:

$$\frac{\partial S}{\partial \mu} = 0 = -2\sum_{i=1}^{N}(x_i-\mu) \quad \Longrightarrow \quad \boxed{\mu_{\rm MLE} = \frac{1}{N}\sum_{i=1}^{N} x_i} \quad (147)$$

¶ **Robust-satisficing:**

- Form of the pdf is not certain.
- $\widetilde{p}(x|\lambda)$ is most reasonable choice of the form of the pdf. We will estimate λ.
- Actual form of the pdf is unknown.
- We wish to choose those parameters to:
  - *Satisfice* the likelihood.
  - To be *robust* to the info-gaps in the shape of the actual pdf which generated the data, or which might generate data in the future.

¶ **Info-gap model:**

$$\mathcal{U}(h,\widetilde{p}) = \left\{p(x): \ p(x)\in\mathcal{P}, \ |p(x)-\widetilde{p}(x|\lambda)| \le h\psi(x)\right\}, \quad h\ge 0 \quad (148)$$

- $\mathcal{P}$ is the set of all normalized and non-negative pdfs on the domain of x.
- $\psi(x)$ is the known envelope function. E.g. $\psi(x)=1$, implying severe uncertainty on tail.
- h is the unknown horizon of uncertainty.

¶ **Question:**

Given the random sample X, and the info-gap model $\mathcal{U}(h,\widetilde{p})$, how should we choose the parameters of the nominal pdf $\widetilde{p}(x|\lambda)$?

¶ **Robustness:**

$$\widehat{h}(\lambda, L_{\rm c}) = \max\left\{h: \ \left(\min_{p\in\mathcal{U}(h,\widetilde{p})} L(p)\right) \ge L_{\rm c}\right\} \quad (149)$$

¶ $m(h) =$ **inner minimum** in eq.(149).

For the info-gap model in eq.(148) $m(h)$ is obtained for the following choices of the pdf at the data points X:

$$p(x_i) = \begin{cases} \widetilde{p}(x_i) - h\psi(x_i) & \text{if } h \le \widetilde{p}(x_i)/\psi(x_i) \\ 0 & \text{else} \end{cases} \quad (150)$$

Choose $p(x) = \widetilde{p}(x)$ for all other x's.
Define:

$$h_{\max} = \min_i \frac{\widetilde{p}(x_i)}{\psi(x_i)} \tag{151}$$

Since $m(h)$ is the product of the densities in eq.(150) we find:

$$m(h) = \begin{cases} \prod_{i=1}^{N} [\widetilde{p}(x_i) - h\psi(x_i)] & \text{if } h \le h_{\max} \\ 0 & \text{else} \end{cases} \tag{152}$$

¶ $m(h)$ **and** $\widehat{h}(\lambda, L_{\rm c})$**:**

- Robustness is the max h at which $m(h) \ge L_{\rm c}$.
- $m(h)$ strictly decreases as h increases.
- Hence robustness is the solution of $m(h) = L_{\rm c}$.
- Hence $m(h)$ is the inverse of $\widehat{h}(\lambda, L_{\rm c})$:

$$m(h) = L_{\rm c} \quad \text{implies} \quad \widehat{h}(\lambda, L_{\rm c}) = h \tag{153}$$

- Plot of $m(h)$ vs. h is plot of $L_{\rm c}$ vs. $\widehat{h}(\lambda, L_{\rm c})$.

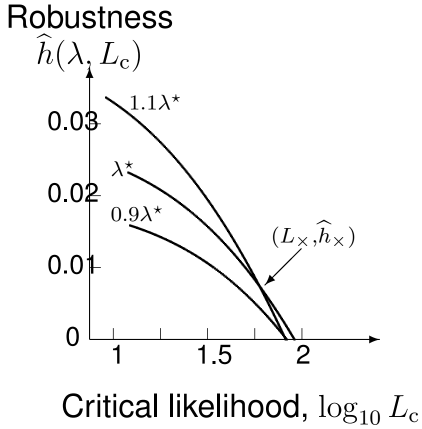


Figure 17: Robustness curves. $\lambda^\star = 3.4065$.

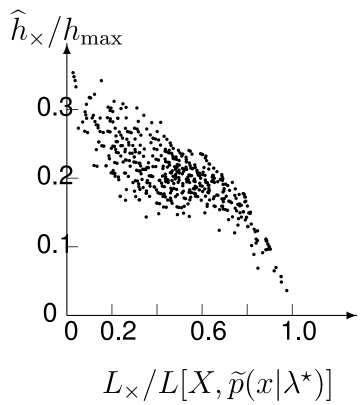


Figure 18: Loci of intersection of robustness curves $\widehat{h}(\lambda^\star, L_c)$ and $\widehat{h}(1.1\lambda^\star, L_c)$.

¶ **Robustness curves** in fig. 17 based on:
- Eqs.(152) and (153).
- Nominal pdf is exponential, $\widetilde{p}(x|\lambda) = \lambda \exp(-\lambda x)$ with $\lambda = 3$.
- Envelope function is constant, $\psi(x) = 1$. Note severe uncertainty on the tail.
- Random sample, X, with $N = 20$.
- MLE of λ, eq.(137): $\lambda^\star = 1/\overline{x}$ where $\overline{x} = (1/N)\sum_{i=1}^N x_i$ is the sample mean.
- Robustness curves for 3 λ's: $0.9\lambda^\star$, $\lambda^\star$, and $1.1\lambda^\star$.

¶ **Robustness of the estimated likelihood is zero for any λ:**
- Likelihood function for λ is $L[\widetilde{p}(x|\lambda)]$.
- Each curve in fig.17, $\widehat{h}(\lambda, L_c)$ vs. L_c, hits horizontal axis when L_c = likelihood:

$$\widehat{h}(\lambda, L_c) = 0 \quad \text{if} \quad L_c = L[\widetilde{p}(x|\lambda)] \tag{154}$$

- $\lambda^\star$ is the MLE of λ. Thus $\widehat{h}(\lambda^\star, L_c)$ hits horizontal axis to the right of $\widehat{h}(\lambda, L_c)$.

¶ **Preferences between estimates of λ:**
- $\widehat{h}(\lambda^\star, L_c) > \widehat{h}(0.9\lambda^\star, L_c) \implies \lambda^\star \succ 0.9\lambda^\star$.
- $\widehat{h}(\lambda^\star, L_c)$ and $\widehat{h}(1.1\lambda^\star, L_c)$ cross at $(L_\times, \widehat{h}_\times)$:
  - $\lambda^\star \succ 1.1\lambda^\star$ for $L_c > L_\times$ and $h < h_\times$.
  - $1.1\lambda^\star \succ \lambda^\star$ else.

¶ 500 repetitions:

- $\lambda^{\star}$ dominates $0.9\lambda^{\star}$.
- Preferences reverse between $\lambda^{\star}$ and $1.1\lambda^{\star}$.
- Normalized $(h_{\times},\ L_{\times})$ in fig. 18.
- Center of cloud: (0.5, 0.2). Typical cross of robustness curves at:
  - ◦ $L_{\rm c}$ about half of best-estimated value.
  - ◦ $\widehat{h}$ about 20% of maximum robustness.

¶ Past and future data-generating processes:

- Data in this example generated from exponential distribution.
- Nothing in data to suggest that exponential distribution is wrong.
- Motivation for info-gap model, eq.(148), is that,
  - ◦ while the *past* has been exponential,
  - ◦ the *future* may not be.

• The robust-satisficing estimate of λ accounts not only for the historical evidence (the sample X) but also for the future uncertainty about relevant family of distributions.

5 Forecasting

¶ Source material: Yakov Ben-Haim, 2009, Info-gap forecasting and the advantage of sub-optimal models, *European Journal of Operational Research,* 197: 203–213.
Forecasting occurs in many domains:

1. Future trajectory of a guided missile, based on measurements of its past trajectory.
2. Lifetime of component or system based on observed performance and wear.
3. Medical prognosis from clinical data.
4. Economic forecast of market performance based on market data.

5.1 Preliminary Example: European Central Bank Interest Rates

Date	Interest rate	Implied λ
1 Jan 1999	4.50	
9 Apr 1999	3.50	0.778
5 Nov 1999	4.00	1.143
4 Feb 2000	4.25	1.063
17 Mar 2000	4.50	1.059
28 Apr 2000	4.75	1.056
9 Jun 2000	5.25	1.105
28 Jun 2000	5.25	1.000
1 Sep 2000	5.50	1.048
6 Oct 2000	5.75	1.045
11 May 2001	5.50	0.957
31 Aug 2001	5.25	0.955

Table 1: Interest rates for overnight loans at the European Central Bank (marginal lending facility). Source: http://www.ecb.int/stats/monetary/rates/html/index.en.html

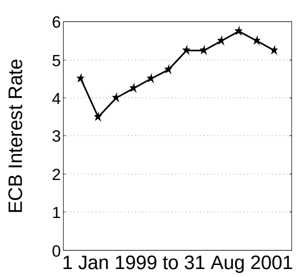


Figure 19: ECB Interest Rates

¶ **ECB overnight interest rates: table 1.**
- First loans: 1999.
- Data through August 2001.

¶ **El-Qaeda attacks in US: 11 Sept 2001.**
- Predict next ECB interest rate?
- **Asymmetric uncertainty:** rate will go down.

¶ **Military force build-up.**
- The adversary is increasing his intel capability at the expense of land-maneuver forces.
- Will this trend continue?

¶ **Questions:**
- How to forecast the rate?
- How to assess confidence in the forecast?

5.2 Info-Gap Forecasting: Formulation

5.2.1 The Estimated System and its Uncertainty

¶ N**-dimensional system** whose average behavior is:

$$y_t = A_t y_{t-1} \tag{155}$$

Zero-mean, additive, random disturbances are ignored and all other inputs are incorporated in the multi-dimensional state vector y_t.

¶ **Solution of eq.(155):**

$$y_{T+k} = \left(\prod_{i=1}^{k} A_{T+i} \right) y_T \tag{156}$$

where the product operator is lefthand matrix multiplication: $\prod_{i=1}^{k} A_{T+i} = A_{T+k} \prod_{i=1}^{k-1} A_{T+i}$.

¶ **1-Step and** k**-Step Forecast:**

- From eq.(156): a k-step process is a 1-step process with coefficient matrix $A^{(k)} = \prod_{i=1}^{k} A_{T+i}$.
- If the matrices A_{T+i} belong to an info-gap model, $\mathcal{U}(h, \widetilde{A})$, then the product matrix $A^{(k)}$ also belongs to an info-gap model, $\mathcal{U}_k(h, \widetilde{A}^k)$:

$$\mathcal{U}_k(h, \widetilde{A}^k) = \left\{ A = \prod_{i=1}^{k} A_i : \ A_i \in \mathcal{U}(h, \widetilde{A}) \right\}, \quad h \geq 0 \tag{157}$$

- Many conclusions about 1-step forecasts hold for k-step forecasts also.

¶ **Info-gap uncertainty** in transition matrices A_t. E.g.:

$$\mathcal{U}(h, \widetilde{A}) = \left\{ A_t, t > T : \ \widetilde{A}_{ij} - h v_{ij} \leq [A_t]_{ij} \leq \widetilde{A}_{ij} + h w_{ij}, \ i, j = 1, \ldots, N \right\}, \quad h \geq 0 \tag{158}$$

- Note asymmetric uncertainty if $v_{ij} \neq w_{ij}$.
- Note constant nominal transition matrix $\widetilde{A}$.

5.2.2 Forecasting with Slope Adjustment

¶ "Slope-adjusted" predictor:

$$y_t^{\mathrm{s}} = B y_{t-1}^{\mathrm{s}} \tag{159}$$

where B is a constant matrix which we are free to choose. The question is: how to choose B?

¶ Vector of average forecast errors for time $t = T + k$ (ignoring zero-mean, additive, random disturbances), based on knowledge of y_T, is:

$$\eta_k(B, A_t) = y_{T+k}^{\mathrm{s}} - y_{T+k} = \left(B^k - \prod_{i=1}^{k} A_{T+i} \right) y_T \tag{160}$$

- Should we really choose $B \neq \widetilde{A}$?
- Judicious choice of B can reliably compensate for deviation of A_{T+i} from $\widetilde{A}$.

5.2.3 Definition of the Robustness Function

¶ Requirement: satisfice the forecast error of mth element at time step k:

$$|\eta_{k,m}(B, A_t)| \le \varepsilon_{\mathrm{c}} \tag{161}$$

¶ Robustness:

$$\widehat{h}(B, \varepsilon_{\mathrm{c}}) = \max \left\{ h : \left(\max_{\substack{A_{T+i} \in \mathcal{U}(h, \widetilde{A}) \\ i=1, \ldots, k}} |\eta_{k,m}(B, A_t)| \right) \le \varepsilon_{\mathrm{c}} \right\} \tag{162}$$

5.2.4 Evaluating the Robustness Function

¶ We evaluate the robustness for 1-step forecast.

¶ The robustness in eq.(162) can be written:

$$\widehat{h}(B, \varepsilon_{\mathrm{c}}) = \max \left\{ h : \left(\max_{A_{T+1} \in \mathcal{U}(h, \widetilde{A})} \eta_{1,m}(B, A_{T+1}) \right) \le \varepsilon_{\mathrm{c}} \right.$$
$$\left. \text{and} \left(\min_{A_{T+1} \in \mathcal{U}(h, \widetilde{A})} \eta_{1,m}(B, A_{T+1}) \right) \ge -\varepsilon_{\mathrm{c}} \right\} \tag{163}$$

¶ The 1-step forecast error for the mth state variable, from eq.(160), is:

$$\eta_{1,m}(B, A_{T+1}) = \underbrace{\sum_{n=1}^{N} [B - \widetilde{A}]_{mn} y_{T,n}}_{\delta} - \sum_{n=1}^{N} [A_{T+1} - \widetilde{A}]_{mn} y_{T,n} \tag{164}$$

δ can be positive or negative and is controlled through the choice of the forecast matrix B.

¶ Define:

$$\theta_c(h) = \max_{A_{T+1} \in \mathcal{U}(h, \widetilde{A})} \sum_{n=1}^{N} [A_{T+1} - \widetilde{A}]_{mn} y_{T,n} \quad (165)$$

$$\theta_a(h) = - \min_{A_{T+1} \in \mathcal{U}(h, \widetilde{A})} \sum_{n=1}^{N} [A_{T+1} - \widetilde{A}]_{mn} y_{T,n} \quad (166)$$

- Contraction axiom implies that $\theta_a(0) = \theta_c(0) = 0$.
- Nesting axiom then implies that $\theta_a(h) \geq 0$ and $\theta_c(h) \geq 0$ and monotonic for all $h \geq 0$.
- Note:

$$\max_h \eta_{1,m} = \delta + \theta_a(h) \leq \varepsilon_c \quad (167)$$

$$\min_h \eta_{1,m} = \delta - \theta_c(h) \geq -\varepsilon_c \iff -\delta + \theta_c(h) \leq \varepsilon_c \quad (168)$$

¶ From eqs.(164)–(166), the robustness is:

$$\widehat{h}(B, \varepsilon_c) = \max \left\{ h : \ \delta + \theta_a(h) \leq \varepsilon_c \ \text{ and } \ -\delta + \theta_c(h) \leq \varepsilon_c \right\} \quad (169)$$

¶ Plotting the robustness.

- Define:

$$\varepsilon(h) = \max \left\{ \delta + \theta_a(h), \ -\delta + \theta_c(h) \right\} \quad (170)$$

- $\varepsilon(h)$ is the inverse of $\widehat{h}(B, \varepsilon_c)$:
  Plot of h vertically vs. $\varepsilon(h)$ horizontally is the same as a
  plot of $\widehat{h}(B, \varepsilon_c)$ vertically vs. ε_c horizontally as in fig. 20.
- Fig. 20: The vertical axis is h or $\widehat{h}(B, \varepsilon_c)$, while the horizontal axis is $\varepsilon(h)$ or ε_c.

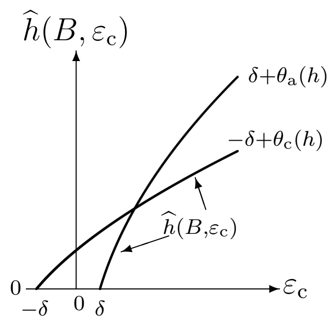


Figure 20: Robustness function based on eqs.(169) and (170).

- The discontinuous slope of $\widehat{h}$ vs ε_c can result in:
  - Crossing robustness curves for different choices of B.
  - Preference for $B \neq \widetilde{A}$.
  - Recall that δ is controlled by the choice of B.

5.2.5 Crossing of Robustness Curves and the Advantage of Sub-Optimal Models

¶ **1-step forecast error,** eq.(164):

$$\eta_{1,m}(B, A_{T+1}) = \underbrace{\sum_{n=1}^{N} [B - \widetilde{A}]_{mn} y_{T,n}}_{\delta} - \sum_{n=1}^{N} [A_{T+1} - \widetilde{A}]_{mn} y_{T,n} \tag{171}$$

¶ **Applies also to k-step error,** with notational change.

¶ **If A_t will be constant at $\widetilde{A}$ in the future,** then the k-step prediction error for the mth state variable is:

$$\eta_{k,m}(B, \widetilde{A}) = \underbrace{\sum_{n=1}^{N} \left(B^k - \widetilde{A}^k \right)_{mn} y_{T,n}}_{\varepsilon^\star} \tag{172}$$

• One is tempted to choose $B = \widetilde{A}$ in order to minimize the anticipated prediction error $\varepsilon^\star$.

• Is this a good choice?

¶ **Theorem:** There exist sub-optimal models for 1-step forecasting which are more robust than optimal models.

5.3 Example: 1-Dimensional System

¶ **The system.** Consider a scalar system whose average behavior evolves as:

$$y_t = \lambda_t y_{t-1} \tag{173}$$

E.g. y_t is a mechanical deflection that increases over time due to decreasing stiffness.

¶ **Asymmetric uncertainty:** λ_t tends to drift up.

$$\mathcal{U}(h, \widetilde{\lambda}) = \left\{ \lambda_t, t > T : \ 0 \le \frac{\lambda_t - \widetilde{\lambda}}{\widetilde{\lambda}} \le h \right\}, \quad h \ge 0 \tag{174}$$

¶ **Slope-adjusted forecaster.**

$$y_t^{\rm s} = \ell y_{t-1}^{\rm s} \tag{175}$$

¶ **Robustness of k-step forecast** with growth coefficient ℓ, defined in eq.(162):

$$\widehat{h}(\ell, \varepsilon_{\rm c}) = \begin{cases} 0 & \text{if } \varepsilon_{\rm c} \le (\ell^k - \widetilde{\lambda}^k) y_T \\ \left(\dfrac{\varepsilon_{\rm c} + \ell^k y_T}{\widetilde{\lambda}^k y_T} \right)^{1/k} - 1 & \text{else} \end{cases} \tag{176}$$

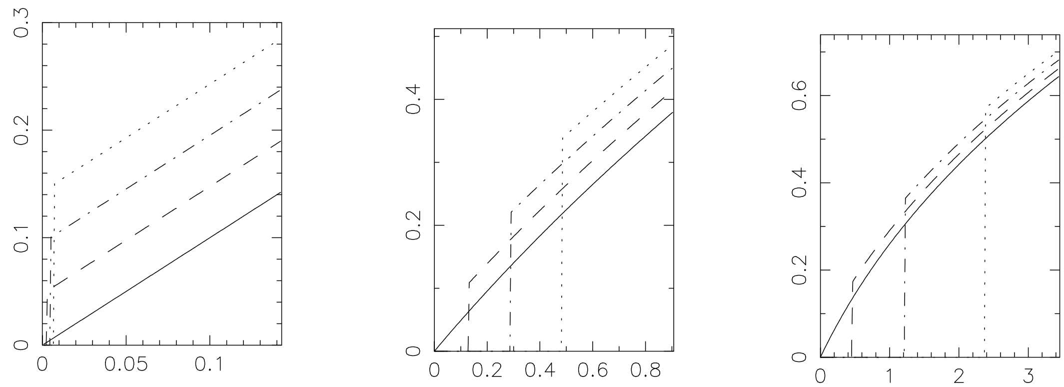


Figure 21: Robustness, $\widehat{h}(\ell, \varepsilon_{\rm c})$, vs. normalized forecast error, $\varepsilon_{\rm c}/\widetilde{\lambda}^k y_T$ for $k = 1,\, 2,\, 3$ from left to right, eq.(176), for $\ell =$ 1.05, 1.1, 1.15, 1.2 from bottom to top curve. $\widetilde{\lambda} = 1.05$, $y_T = 1$.

¶ **Numerical example,** fig. 21.

- Lowest curve in each frame is nominal forecaster: $\ell = \widetilde{\lambda} = 1.05$.
- ℓ increases by 0.05 with each higher curve.
- Horizontal axis: satisficed forecast error, $\varepsilon_{\rm c}$, normalized to nominal forecast value, $\widetilde{\lambda}^k y_T$.
- **1-step forecast** (left frame):

◦ Slope-adjusted predictors are far more robust than the nominal predictor for essentially all levels of forecast error $\varepsilon_{\rm c}$.

◦ For instance, consider 5% fractional forecast error, $\varepsilon_{\rm c}/\widetilde{\lambda}^k y_T = 0.05$.
$\widehat{h}(1.05, \varepsilon_{\rm c}) = 0.050$ (bottom curve), and $\widehat{h}(1.2, \varepsilon_{\rm c}) = 0.19$ (top curve).
The slope-adjusted predictor is about 4 times more robust than the nominal predictor.

- **2- and 3-step forecast** (middle and right frames):

◦ robustness premium of slope-adjusted forecaster, $\ell > \widetilde{\lambda}$, compared to the nominal predictor, $\ell = \widetilde{\lambda}$, becomes smaller as the horizon of the prediction increases.

5.4 Robustness and Probability of Forecast Success

¶ **1-step forecast error** of m variable, from eq.(160), is:

$$\eta_{1,m}(B, A_{T+1}) = \sum_{n=1}^{N} [B - A_{T+1}]_{mn}\, y_{T,n} \tag{177}$$

¶ **Forecast is successful if:**

$$|\eta_{1,m}(B, A_{T+1})| \le \varepsilon_{\mathrm{c}} \tag{178}$$

• This can be written explicitly as:

$$-\varepsilon_{\mathrm{c}} + \sum_{n=1}^{N} [B - \widetilde{A}]_{mn} y_{T,n} \le \underbrace{\sum_{n=1}^{N} \left[A_{T+1} - \widetilde{A}\right]_{mn} y_{T,n}}_{u} \le \varepsilon_{\mathrm{c}} + \sum_{n=1}^{N} [B - \widetilde{A}]_{mn} y_{T,n} \tag{179}$$

which defines the variable u.

• Recalling the definition of δ in eq.(164), the condition for forecast success in eq.(179) becomes:

$$\delta - \varepsilon_{\mathrm{c}} \le u \le \delta + \varepsilon_{\mathrm{c}} \tag{180}$$

¶ **Probability of forecast success:**

• $F(u)$ is unknown cumulative probability distribution of u.

• Probability of forecast success with model B:

$$P_{\mathrm{s}}(B) = F(\delta + \varepsilon_{\mathrm{c}}) - F(\delta - \varepsilon_{\mathrm{c}}) \tag{181}$$

¶ **Is robustness, $\widehat{h}(B, \varepsilon_{\mathrm{c}})$, a proxy for probability of success, $P_{\mathrm{s}}(B)$?**

Yes, in a wide range of situations.