Structural Equation Models with Latent Variables

Francisco Peñagaricano
University of Florida

Causal Effects

Decipher causal relationships is the ultimate goal in most studies involving complex traits

▪unravel causal relations among variables can be used to predict the behavior of complex systems

Inferring causal effects from observational data is difficult due to the presence of potential confounders

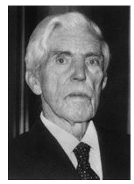
▪suitable methodologies, e.g. **structural equation models** are already available and have been used in other fields

Structural Equation Models

modeling of causal relationships between multiple traits

Causality and Structural Equation Models

Wright



Haavelmo

the founding fathers considered SEM as a mathematical tool for drawing causal conclusions from a combination of observational data and theoretical assumptions

Structural Equation Models

modeling of causal relationships between multiple traits

Causality and Structural Equation Models

Structural Equation Modeling is an inference tool:

INPUTS:

- **qualitative causal assumptions**
- **empirical data**

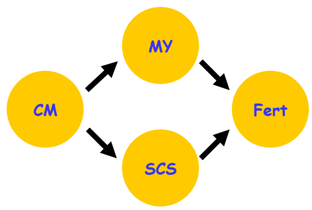
OUTPUTS:

- **quantitative causal conclusions**
- **statistical measures of the fit of the causal model**

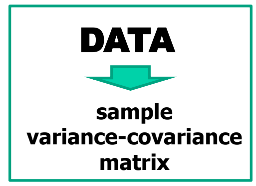
Path Analysis

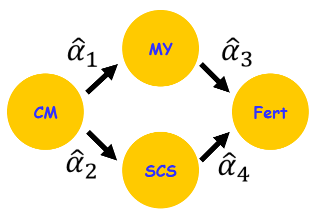
(SEM considering only **observed variables**)

qualitative causal assumptions



empirical data





quantitative causal conclusions
(path coefficients)

fit of the causal model
(χ^2 value)

Structural Equation Models

- one of the most remarkable features of SEM is the ability to consider **latent variables**

Latent variables: variables that are **not observable** or are **not directly measurable**, but are **characterized** in the model from several **observed variables**

Latent variables

variables that are **not observable** or are **not directly measurable**, but are **characterized** in the model from several **observed variables**

Latent variables

variables that are **not observable** or are **not directly measurable**, but are **characterized** in the model from several **observed variables**

- **Intelligence**
- **Meat quality**
- **Fertility**

Latent variables

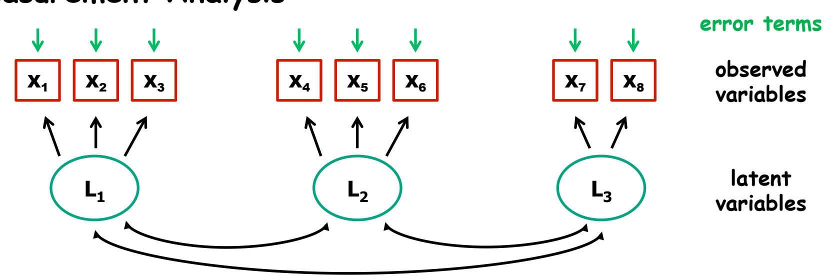
variables that are **not observable** or are **not directly measurable**, but are **characterized** in the model from several **observed variables**

latent variables allow modeling complex phenomena reducing at the same time the dimensionality of the data

- many phenotypes can be combined in a model to represent an underlying concept of interest

Structural Equation Modeling

Measurement Analysis

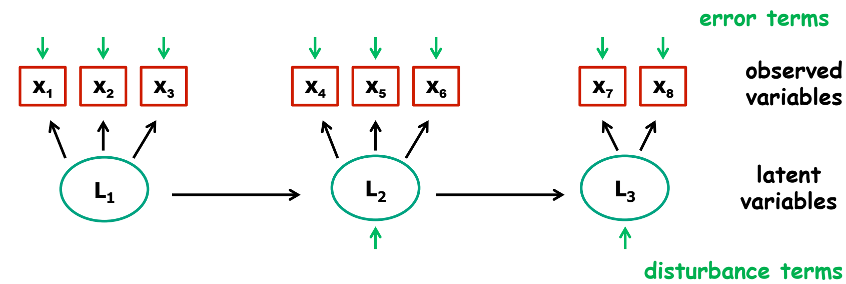


Confirmatory Factor Analysis

- Measure latent variables
- Reduce the dimensionality of the data
- Test the statistical significance of factor loadings
- Precursor of a hybrid model

Structural Equation Modeling

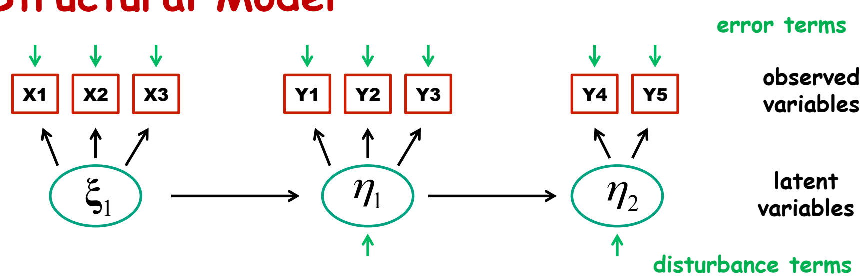
Structural Analysis



Hybrid Model

- Evaluate presumed causal relations among latent variables

Structural Model



measurement model

$$\mathbf{x} = \mathbf{\Lambda}_x \boldsymbol{\xi} + \boldsymbol{\delta}$$

$$\mathbf{y} = \mathbf{\Lambda}_y \boldsymbol{\eta} + \boldsymbol{\varepsilon}$$

structural model

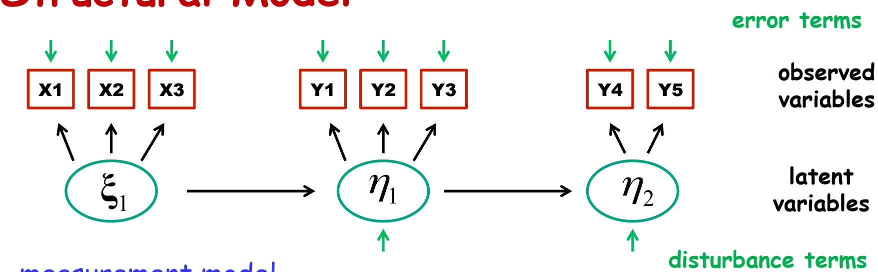
$$\boldsymbol{\eta} = \mathbf{B}\boldsymbol{\eta} + \mathbf{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta}$$

Assumptions:

$$E(\boldsymbol{\zeta}) = E(\boldsymbol{\delta}) = E(\boldsymbol{\varepsilon}) = 0$$

$$Cov(\xi,\zeta) = Cov(\xi,\delta) = Cov(\xi,\varepsilon) = Cov(\delta,\varepsilon) = Cov(\zeta,\delta) = Cov(\zeta,\varepsilon) = 0$$

Structural Model



measurement model

$$\mathbf{x} = \mathbf{\Lambda}_x \boldsymbol{\xi} + \boldsymbol{\delta}$$

$$\mathbf{y} = \mathbf{\Lambda}_y \boldsymbol{\eta} + \boldsymbol{\varepsilon}$$

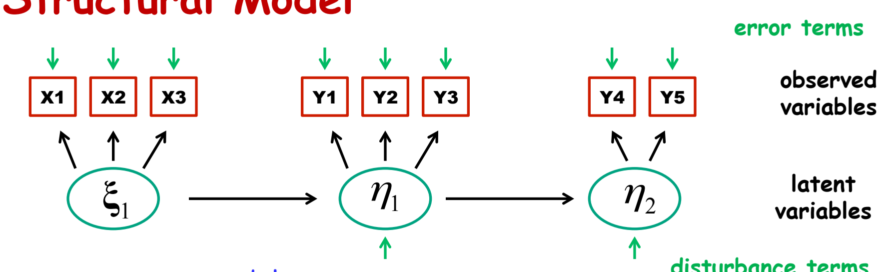
structural model

$$\boldsymbol{\eta} = \mathbf{B}\boldsymbol{\eta} + \mathbf{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta}$$

$$\mathbf{\Lambda}_x = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{bmatrix} \quad \mathbf{\Lambda}_y = \begin{bmatrix} \lambda_4 & 0 \\ \lambda_5 & 0 \\ \lambda_6 & 0 \\ 0 & \lambda_7 \\ 0 & \lambda_8 \end{bmatrix} \quad \mathbf{\Gamma} = \begin{bmatrix} \gamma_{11} \\ 0 \end{bmatrix} \quad \mathbf{B} = \begin{bmatrix} 0 & 0 \\ \beta_{21} & 0 \end{bmatrix}$$

Parameters

Structural Model



measurement model

$$\mathbf{x} = \mathbf{\Lambda}_x \boldsymbol{\xi} + \boldsymbol{\delta}$$

$$\mathbf{y} = \mathbf{\Lambda}_y \boldsymbol{\eta} + \boldsymbol{\varepsilon}$$

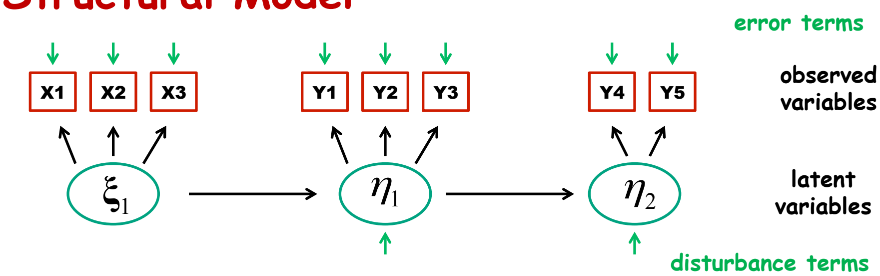
structural model

$$\boldsymbol{\eta} = \mathbf{B}\boldsymbol{\eta} + \mathbf{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta}$$

$$\mathbf{\Theta}_\delta = \begin{bmatrix} \theta_{11}^{(\delta)} & & \\ 0 & \theta_{22}^{(\delta)} & \\ 0 & 0 & \theta_{33}^{(\delta)} \end{bmatrix} \quad \mathbf{\Theta}_\varepsilon = \begin{bmatrix} \theta_{11}^{(\varepsilon)} & & & & \\ 0 & \theta_{22}^{(\varepsilon)} & & & \\ 0 & 0 & \theta_{33}^{(\varepsilon)} & & \\ 0 & 0 & 0 & \theta_{44}^{(\varepsilon)} & \\ 0 & 0 & 0 & 0 & \theta_{55}^{(\varepsilon)} \end{bmatrix} \quad \mathbf{\Psi} = \begin{bmatrix} \psi_{11} & \\ 0 & \psi_{22} \end{bmatrix} \quad \mathbf{\Phi} = \begin{bmatrix} \phi_{11} \end{bmatrix}$$

VCOV matrices

Structural Model



Model Estimation

- **VCOV matrices** are the fundamental units of the analysis

Estimation: find parameters that reproduce as close as possible the **observed VCOV matrix**

Structural Equation Modeling

Σ	variance-covariance matrix for the entire population
S	variance-covariance matrix computed from a sample of the population
$\hat{\Sigma}$	model-based (fitted) variance-covariance matrix
$S - \hat{\Sigma}$	residual variance-covariance matrix

Estimation:

- find parameters that make $\hat{\Sigma}$ as close as possible to S
- **Minimize a fitting function** $F(S, \hat{\Sigma})$

Structural Equation Modeling

$\boldsymbol{\Sigma}$	variance-covariance matrix for the entire population
$\mathbf{S}$	variance-covariance matrix computed from a sample of the population
$\hat{\boldsymbol{\Sigma}}$	model-based (fitted) variance-covariance matrix
$\mathbf{S}-\hat{\boldsymbol{\Sigma}}$	residual variance-covariance matrix

- **Minimize a fitting function $F(S,\hat{\Sigma})$**

if variables follow a multivariate normal distribution, the **ML estimates** are those that minimize the following fitting function:

$$F_{ML} = \log\left|\hat{\boldsymbol{\Sigma}}\right| + tr\left(\mathbf{S}\times\hat{\boldsymbol{\Sigma}}^{-1}\right) - \log|\mathbf{S}| - (p+q)$$

Structural Equation Modeling

$$F_{ML} = \log\left|\hat{\boldsymbol{\Sigma}}\right| + tr\left(\mathbf{S}\times\hat{\boldsymbol{\Sigma}}^{-1}\right) - \log|\mathbf{S}| - (p+q)$$

test for model fit

$$(N-1)\cdot F_{ML} \sim \chi^2_{df}$$

df: difference between the number of unique elements in the VCOV matrix and the number of free parameters in the model

In SEM: **we want high P-values**

(we want to NO reject H_0)

Structural Equation Modeling

$$F_{ML} = \log\left|\hat{\mathbf{\Sigma}}\right| + tr\left(\mathbf{S}\times\hat{\mathbf{\Sigma}}^{-1}\right) - \log\left|\mathbf{S}\right| - \left(p+q\right)$$

test for model fit

$$\left(N-1\right)\cdot F_{ML} \sim \chi^2_{df}$$

- **global test:** it evaluates simultaneously all the restrictions imposed in the variance-covariance matrix
- if the **test is significant**, i.e. we reject H_0, **the source of the lack of fit is unclear**
- depends on the **sample size** and the **number of parameters**

Structural Equation Modeling

Alternatives fit indices

- **Standardized Root Mean Square Residual (SRMR)**

$$SRMR = [(p+q)^{-1}(\mathbf{e}'\mathbf{W}\mathbf{e})]^{1/2}$$

SRMR ≤ 0.08

- **Root mean square error of approximation (RMSEA)**

$$RMSEA = \sqrt{\frac{\hat{\lambda}_N}{df}} = \sqrt{\frac{max(\chi^2 - df, 0)}{df(N-1)}}$$

RMSEA ≤ 0.06

- **Tucker-Lewis index (TLI)**

$$TLI = \frac{\chi^2_0/df_0 - \chi^2_k/df_k}{\chi^2_0/df_0 - 1}$$

TLI ≥ 0.95

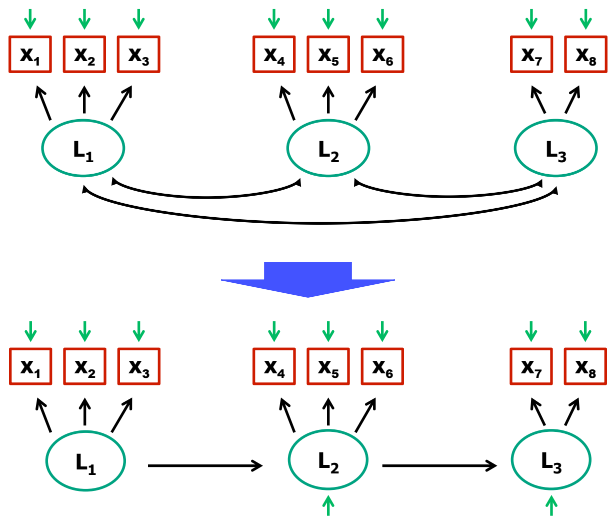
- **Comparative fit index (CFI)**

$$CFI = \frac{max(\chi^2_0 - df_0, 0) - max\left(\chi^2_k - df_k, 0\right)}{max(\chi^2_0 - df_0, 0)}$$

CFI ≥ 0.95

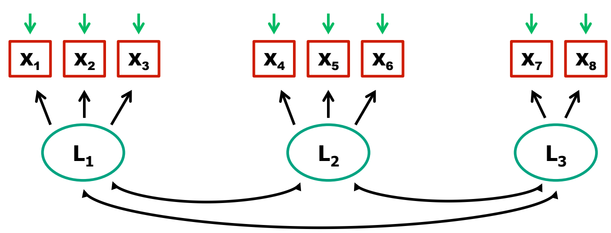
Latent Variable Modeling

in a quantitative genetic context



Latent Variable Modeling

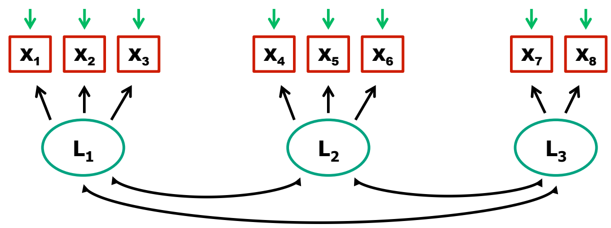
in a quantitative genetic context



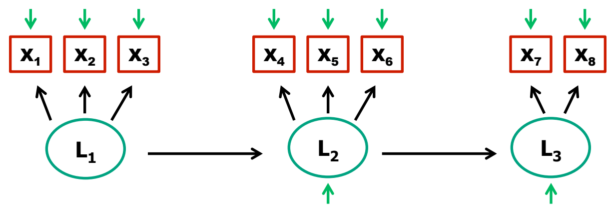
Associations between latent variables can be explained not only by **causal links** between them, but also by **genetic reasons**

Latent Variable Modeling

in a quantitative genetic context

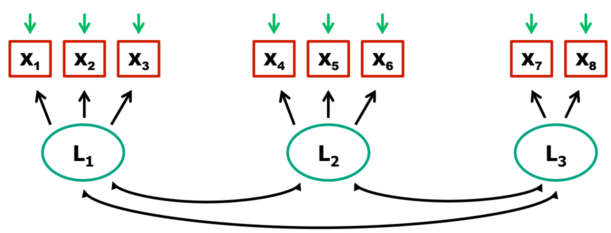


causal links can be masked
by genetic covariances

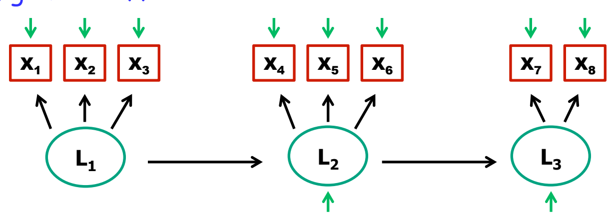


Latent Variable Modeling

in a quantitative genetic context



adjust the data for possible
(confounder) genetic effects



Searching for causal networks involving latent variables in complex traits

The approach can be summarized in the following steps:

1. modeling different latent variables
2. evaluating measurement models using **CFA**
3. fitting a multi-trait animal model using **factor scores**
4. fitting **SEM** for assessing causal links between latent variables

- relevant: evaluate model fit and estimate model parameters

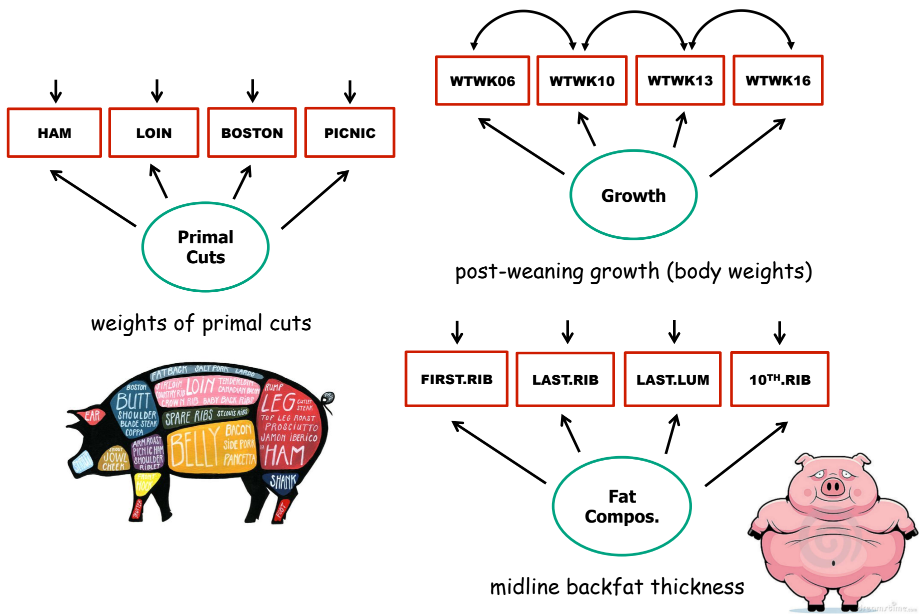
Peñagaricano et al. (2015) J Anim Sci. 93: 4617-4623

Searching for causal networks involving latent variables in complex traits

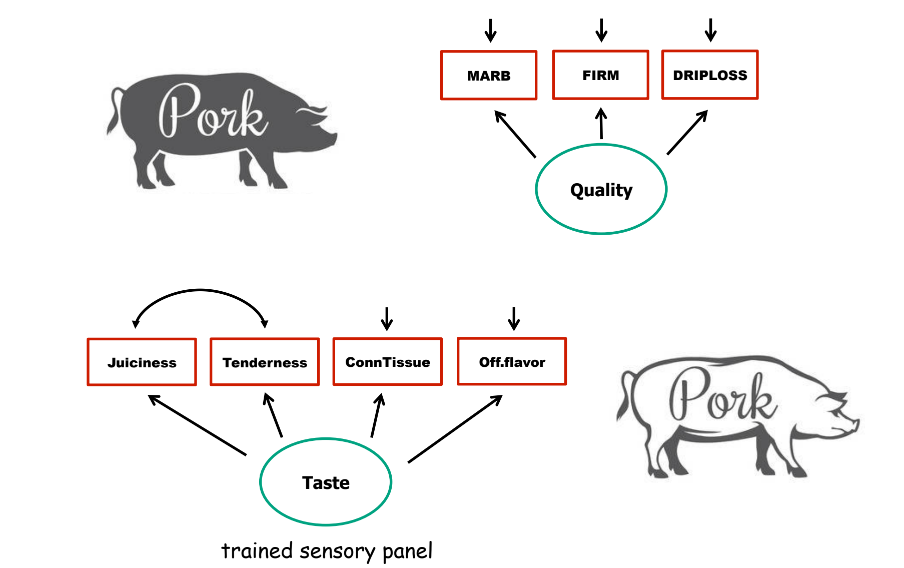
- Application: evaluate possible causal relationships between growth, carcass and meat quality traits in pigs

❖ **phenotypic data:**
dataset with 413 F_2 pigs for which **several phenotypes** were **recorded over time**

Modeling different latent variables

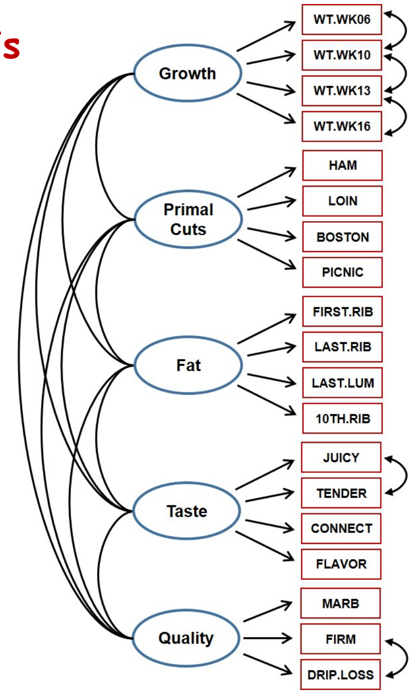


Modeling different latent variables



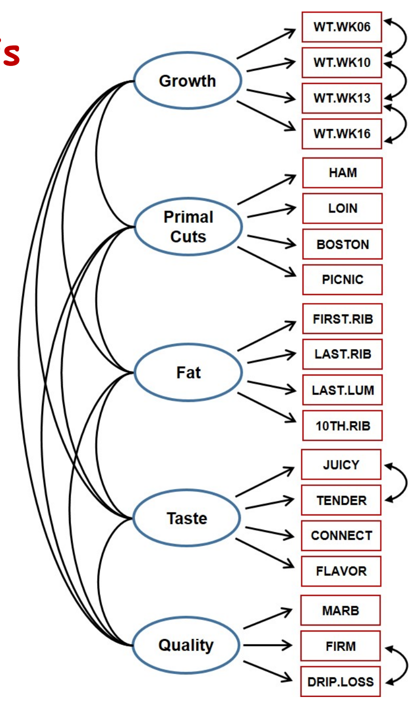
Confirmatory Factor Analysis

- ☑ Model Fit Evaluation
- ☑ Factor Loadings



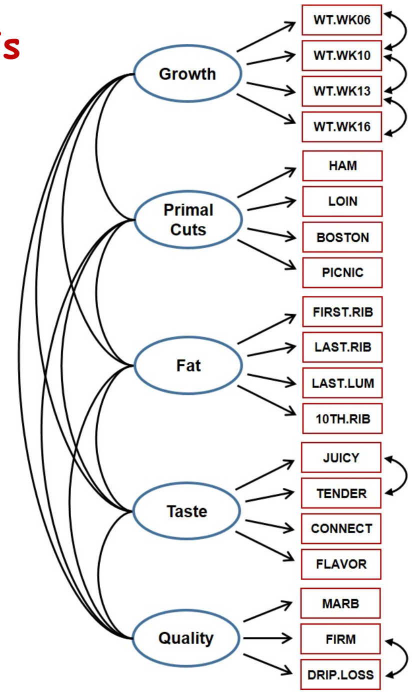
Confirmatory Factor Analysis

Trait	Estimate	SE	Z value	R^2
Growth				
WT.WK06	0.667	0.058	11.50	0.445
WT.WK10	0.779	0.059	13.20	0.607
WT.WK13	0.906	0.054	16.78	0.821
WT.WK16	0.883	0.050	17.66	0.779
Primal Cuts				
HAM	0.865	0.048	18.02	0.748
LOIN	0.792	0.044	18.00	0.627
BOSTON	0.674	0.046	14.65	0.454
PICNIC	0.664	0.045	14.76	0.441
Fat Composition				
FIRST.RIB	0.550	0.048	11.46	0.303
LAST.RIB	0.580	0.067	8.66	0.336
LAST.LUM	0.580	0.047	12.34	0.336
10TH.RIB	0.862	0.060	14.37	0.743
Taste				
JUICY	0.183	0.052	3.52	0.033
TENDER	0.651	0.088	7.40	0.424
CONNECT	0.897	0.115	7.80	0.805
OFF.FLAVOR	−0.171	0.063	−2.71	0.030
Quality				
MARB	0.584	0.090	6.49	0.341
FIRM	0.240	0.077	3.12	0.058
DRIP.LOSS	−0.386	0.069	−5.59	0.149



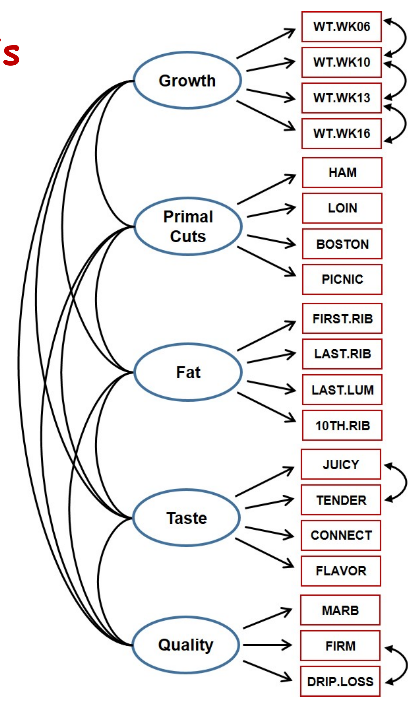
Confirmatory Factor Analysis

define a measurement model for the relationship between **multivariate observations** and **underlying factors**



Confirmatory Factor Analysis

model implied **variance-covariance matrix** of the latent variables !?



Searching for causal networks involving latent variables in complex traits

The approach can be summarized in the following steps:

1. modeling different latent variables
2. evaluating measurement models using **CFA**
3. **fitting a multi-trait animal model using factor scores**

I. estimate **factor scores** $\hat{\xi}_{ij}$

best predictor: conditional expectation of ξ_i given $\mathbf{x_i}$

$$E(\boldsymbol{\xi}_i|\mathbf{x}_i) = E(\boldsymbol{\xi}_i) + Cov(\boldsymbol{\xi}_i, \mathbf{x}_i{}',)Var(\mathbf{x}_i)^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_x) = \hat{\xi}_i$$

Searching for causal networks involving latent variables in complex traits

The approach can be summarized in the following steps:

1. modeling different latent variables
2. evaluating measurement models using **CFA**
3. **fitting a multi-trait animal model using factor scores**

I. estimate **factor scores** $\hat{\xi}_{ij}$

II. fit a **multi-trait animal model** with $y = \hat{\xi}_{ij}$

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{Zu} + \boldsymbol{\varepsilon} \qquad \begin{bmatrix}\mathbf{u}\\ \boldsymbol{\varepsilon}\end{bmatrix} \sim N\left(\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}\mathbf{G_0} \otimes \mathbf{A} & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\Psi}_0 \otimes \mathbf{I_n}\end{bmatrix}\right)$$

Searching for causal networks involving latent variables in complex traits

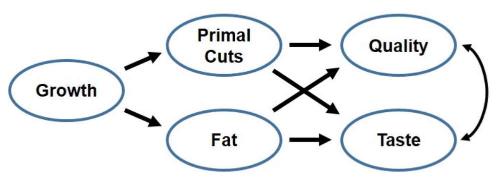
The approach can be summarized in the following steps:

1. modeling different latent variables
2. evaluating measurement models using **CFA**
3. fitting a multi-trait animal model using **factor scores**
4. fitting **SEM** for assessing causal links between latent variables

after adjusting the $\hat{\xi}_{ij}$ for the additive genetic effects,

$$\mathbf{w} = \mathbf{Bw} + \boldsymbol{\Gamma}\mathbf{v} + \boldsymbol{\zeta}$$

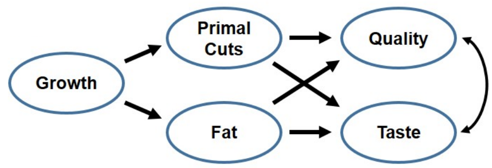
Structural Analysis



☑ **Model Fit Evaluation**

$$\chi^2 = 5.9,\ df = 3,\ Pvalue = 0.110$$

Structural Analysis

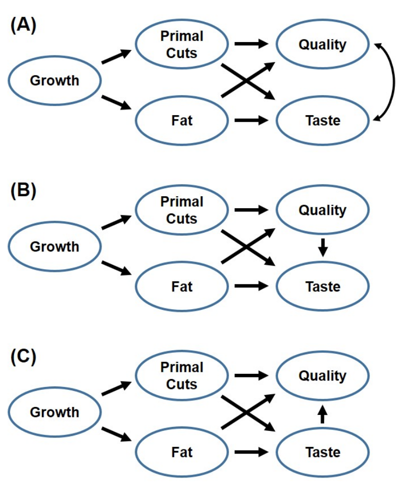


Path	Estimate	SE	Z value
Growth → Cuts	0.514	0.044	11.68
Growth → Fat	0.140	0.056	2.50
Cuts → Quality	−0.500	0.031	16.13
Cuts → Taste	0.095	0.049	1.94
Fat → Quality	0.695	0.029	23.96
Fat → Taste	0.288	0.047	6.13
Quality ↔ Taste	0.140	0.027	5.19
Growth → Quality	−0.160	0.047	3.20

- Primal Cuts has a **negative causal effect** on Quality
- The **total effect** of **Growth on Quality** is equal to $-\mathbf{0.16}$

Statistically Equivalent Models

causal models that can have the same statistical consequences; these models fit any set of data equally well



$\boxed{\chi^2 = 5.9,\ df = 3,\ Pvalue = 0.110,\ AIC = 64.30}$

identical implied VCOV matrices

identical residuals

χ^2 values and goodness-of-fit indices

Statistically Equivalent Models

- translation from a causal hypothesis into a statistical model is **imperfect**

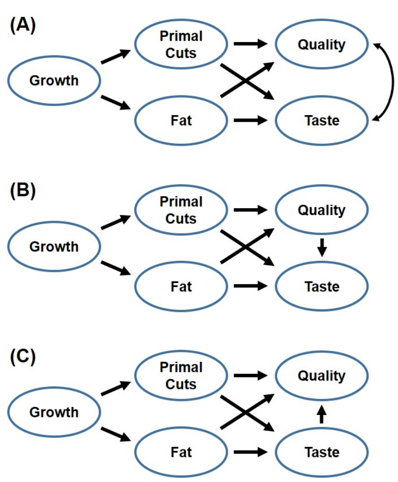
this imperfection arises because:

the asymmetrical links of the causal process are replaced by symmetrical relationships involving probability distributions

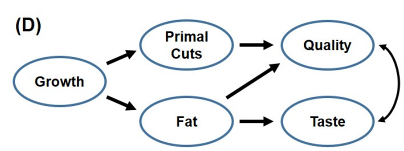
- statistically equivalent models illustrate the **limit** on the ability to infer causality from statistical data alone

Structural Analysis

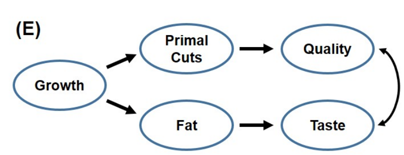
Alternative Causal Structures



$\chi^2 = 5.9, df = 3, Pvalue = 0.110, AIC = 64.30$



$\chi^2 = 10.5, df = 4, Pvalue = 0.014, AIC = 65.91$



$\chi^2 = 245, df = 5, Pvalue \leq 0.001, AIC = 392$