Population Psychometrics

Geographical, temporal and demographic structure

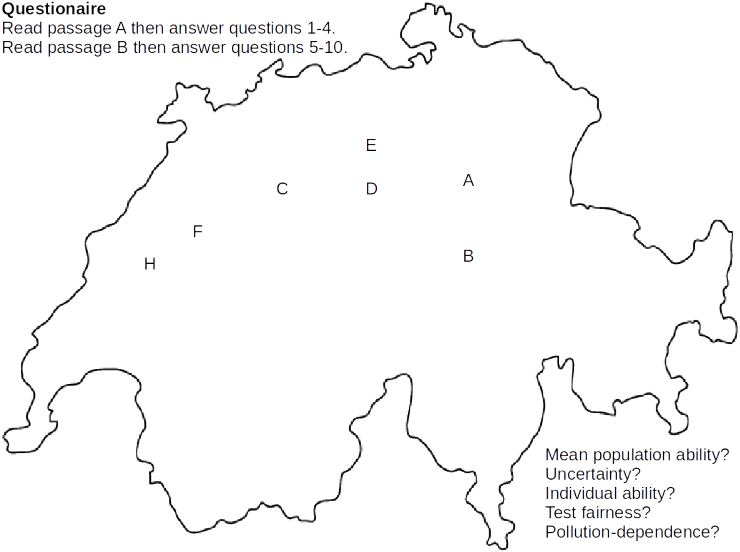
Justin Chumbley[1]

[1]ETH, UZH Zurich

April 27, 2016

Population psychometrics

Questionaire
Read passage A then answer questions 1-4.
Read passage B then answer questions 5-10.



Mean population ability?
Uncertainty?
Individual ability?
Test fairness?
Pollution-dependence?

Population psychometrics

$$
\begin{aligned}
Y_{ijs} &= I(Y^*_{ijs} > 0) \\
Y^*_{ijs} &= \beta_{0js} + \beta_{ijs} + \epsilon_{ijs} \\
\beta_{0js} &= \lambda_{00} + U_{0js} \\
\beta_{ijs} &= \lambda_{i0} \\
U_{0js} &= V_s + S_j
\end{aligned}
$$

- Stein
- Krige
- Simpson
- Pearl
- Fisher

Identifiability constraints, ϵ_{ijs} iid logistic (no LD)
$\beta_{ijs} = \lambda_{i0}$ (no random or fixed DIF)

Population psychometrics

$$\begin{aligned}
\mathrm{Y}_{ijs} &= I(Y^*_{ijs} > 0) \\
\mathrm{Y}^*_{ijs} &= \beta_{0js} + \beta_{ijs} + \epsilon_{ijs} \\
\beta_{0js} &= \lambda_{00} + U_{0js} \\
\beta_{ijs} &= \lambda_{i0} + W_{ijs} \\
\mathrm{U}_{0js} &= V_s + S_j
\end{aligned}$$

- Stein
- Krige
- Simpson
- Pearl
- Fisher

MMSE and regularization
Between-subject variablity in DIF.

Population psychometrics

$$
\begin{aligned}
\mathrm{Y}_{ijs} &= I(Y^*_{ijs} > 0) \\
\mathrm{Y}^*_{ijs} &= \beta_{0js} + \beta_{ijs} + \epsilon_{ijs} \\
\beta_{0js} &= \lambda_{00} + U_{0js} \\
\beta_{ijs} &= \lambda_{i0} + W_{ijs} \\
\mathrm{U}_{0js} &= V(s_j) + S_j
\end{aligned}
$$

- Stein
- Krige
- Simpson
- Pearl
- Fisher

$V(s) \sim Process(\mu, C)$.
Item recommendation.

Population psychometrics

$$
\begin{aligned}
Y_{ijs} &= I(Y^*_{ijs} > 0) \\
Y^*_{ijs} &= \beta_{0js} + \beta_{ijs} + \epsilon_{ijs} \\
\beta_{0js} &= \lambda_{00} + \lambda_{01} X_{js} + U_{0js} \\
\beta_{ijs} &= \lambda_{i0} \\
U_{0js} &= V_k + S_j
\end{aligned}
$$

- Stein
- Krige
- Simpson
- Pearl
- Fisher

"Exogeneity" LD: $U_{0js} \perp\!\!\!\perp X_{js}$, $X_{js} \neq f(V_k)$.
Graphical identifiability criterea/stratified randomization.

Conclusion: Dependence misspecification

- Over-fit λ (underestimated s.e. via local dependence within testlet, subject, region), but applies to spatial/temporal dependence too!
- Biased λ (Simpson)
- Overfit β (relative to MMSE James-Stein)
- Forgo β prediction/generalization!

Population psychometrics: Dependence mis-specification

- Questions

$$
\begin{aligned}
Y_{ijs} &= I(Y^*_{ijs} > 0) \\
Y^*_{ijs} &= \beta_{0js} + \beta_{ijs} + \epsilon_{ijs} \\
\beta_{0js} &= \lambda_0 + \lambda_1 X_{js} + U_{js} \\
\beta_{ijs} &= \beta_i \text{ (no random or fixed DIF)} \\
U_{js} &= V_s + S_j
\end{aligned}
\tag{1}
$$

- Stein, Kridge, Bonferonni, Simpson, Berkson, Pearl, Fisher,...
- Prediction. General index + dependence (beyond exchangeable "clusters": stationarity, isotropy, etc). $\beta_0(s_j) = \lambda_0 + \lambda_1 X(s_j) + U(s_j), s_j \in S$ $U(s) \sim Process(\mu, C)$. Item recommendation!
- Estimation. MMSE biased! (then ridge logistic regression). Adaptive regularization/shrinkage, adaptive control over-fitting, extrema and FWER (c.f. Bonferonni). DIF smoothing/variance component tests. DIF estimation penalty (regularization) vs test penalty. Discrete/continuous DIF index? (constrained minimum MSE: 1. unbiasedness 2. regularization justified by data-generating model. Full

$$
\begin{aligned}
Y_{ijs} &= I(Y^*_{ijs} > 0)\\
Y^*_{ijs} &= \beta_{0js} + \beta_{ijs} + \epsilon_{ijs} \text{ (i.i.d. standard logistic + identifiability)}\\
\beta_{0js} &= \lambda_0 + \lambda_1 X_{js} + U_{js}\\
\beta_{ijs} &= \beta_i \text{ (no random or fixed DIF)}\\
U_{js} &= V_s + S_j
\end{aligned}
\tag{2}
$$

- Mis-model dependence: over-fit λ_0, λ_1 (underestimated s.e. standard HM justification (dependence within-testlet, subject, region), but applies to spatial/temporal dependence too!), bias estimation λ_0, λ_1 (Simpson), overfit β_{0jk} (relative to MMSE James-Stein), forgoes β_{0jk} prediction (generalization to population)!
- Logical dependence
- Recoverability: GPs, Copulas, ...Logistic processes (2015). More general alternative to exchangeability, e.g. temporal fatigue model (c.f. item response drift).

- $\epsilon_i(s)$ a logistic distribution or process: standard, independent? independent of other component processes? stationary? isotropic?

Population Psychometrics

$$\Pr(Y_{p1} = 1|p) = \sum_{y_{p2} \in \{0,1\}} Pr(Y_{p1} = 1, Y_{p2} = y_{p2}|\theta_p) = \frac{exp(\theta_p - \beta_1) + exp(2\theta_p - \beta_1 - \beta_2 + \lambda)}{1 + exp(\theta_p - \beta_1) + exp(\theta_p - \beta_2) + exp(2\theta_p - \beta_1 - \beta_2 + \lambda)}$$

- "Unreproducibility": Margins not logistic (Rasch) if $= 0$ (i.e. locally dependent) and β_1 not "difficulty parameter" (not simply the location on the latent trait for which the probability of responding correctly is 0.5).
- Similar problems also occur in random-effect testlet models (Bradlow, Wainer, Wang, 1999) and other types of conditional models (e.g., the dynamic Rasch model of Verhelst Glas, 1993).
- Marginal models (see e.g., Molenberghs Verbeke, 2005). Univariate margins and the dependence structure are modeled separately.
- Want logistic marginals, no single or generally accepted multivariate logistic distribution (see e.g., Kotz, Balakrishnan, Johnson, 2000), c.f. MVN and multivariate random effect probit models (Ashford Sowden, 1970).
- MVN has the attractive property that its univariate margins

Hidden structure

- IRT non-robust to strong CI ("local independence") assumptions: factorize over items? subjects?
- IRT assumptions to restrict general pattern: distributional/functional form, invariances e.g. conditional independence and stationarity/isotropy
- Independence almost never, conditional independence almost always.
- Expand model until conditional independence (divide and conquer)?
- HM for prediction, smoothing/power-sharing, modular extensibility, interpret new population (hyper)parameters/ latent regressions, but biased, unrecoverable (marginal vs conditional interpretation).
- Dependence typology (avoid general dependency! Qualitative, sparse parametrization of dependency, few effective parameters faced with multiple fitting/testing problem).
- Within-subject variation across $\{Y_i : i \in I\}$: structure/color of noise? testlet redundancy + exchangeable "clusters" de

Population Psychometrics: spatial, temporal and demographic structure

- several possible approaches to model dependencies
- Two typical problems that are associated with some of the most popular models for local item dependencies are non-reproducibility and the impossibility of interpreting β_i as the difficulty of item i. To illustrate these problems, we take the constant combination interaction model (CCI) of Hoskens and De Boeck (1997) as a starting point and for simplicity, only the case with two items (1 and 2) showing residual dependency is considered. $Pr(Y_{p1} = y_{p1}, Y_{p1} = y_{p1}|\theta_p) = \frac{exp(y_{p1}(\theta_p - \beta_1) + y_{p2}(\theta_p - \beta_2) + y_{p1}y_{p2}\lambda)}{1 + exp(\theta_p - \beta_1) + exp(\theta_p - \beta_2) + exp(2\theta_p - \beta_1 - \beta_2 + \lambda)}$ Note that conditional on p , the parameter expresses the conditional log odds ratio for the item pair given the responses on all other items, hence the CCI model only reduces to the familiar Rasch model in case equals zero (see Tuerlinckx De Boeck, 2004, pp.303-304, for the interpretation of higher-order associations between 3 or more items)

Population Psychometrics: marginal vs conditional models

Let Y p = (Y p1 , Y p2 , . . . , Y pI) T represent the (I 1) binary anomaly out- come vector (i.e., with I anomalies) for the pth infant. Assume Y pi is a manifest indicator of the event that some unobserved latent continuous variable X pi (following a logistic distribution with location 0 and scale 1) exceeds a threshold, which can be taken to be zero without loss of generality. The event Y pi = 1 indicates the presence of the ith anomaly in infant p and Y pi = 0 indicates the abscence. Specifically let Y pi = I(X pi ¿ 0), with X pi = i (p i) + pi , such that Pr(Y pi = y pi — p) = Pr(X pi ¿ 0— p) = Pr(pi ¿ i (p i)) = 1 F pi (i (p i)) exp (y pi i (p i)) , = 1 + exp (i (p i)) with p being an infant-specific intercept representing the unobserved severity at which an infant p has been affected (a larger value means that the infant is more severely affected), the parameters i and i rep- resenting the discrimination and severity degree of the ith anomaly, re- spectively. The random variables pi are also called the latent residuals. This model is called the

Population Psychometrics: marginal vs conditional models

- Logistic vector random fields for dependency modeling within and between subjects (testlets and space)? However, research on scalar (univariate) time series or stochastic processes with logistic marginal distributions has been rather minimal compared to the voluminous work that has been carried out on Gaussian time series or stochastic processes; see Mansfield and Hensley (1960), Arnold (1988, 1992, 1993), Arnold and Robertson (1989), Tan and Piantadosi (1991), Sim (1993), Grasman (1998), Newman et al. (2004), Soboleva and Pleasants (2003), and Silva et al. (2005). Very few multiple time series or vector (multivariate) stochastic process with logistic marginals have been discussed in the literature. This motivates us to introduce vector logistic random fields in space and/or time, whose special features are that they allow for any possible mean structure or covariance matrix structure, just as the Gaussian random field does. Suppose that $Z(x) = (Z1(x), \ldots, Zm(x))$, x D is an m-variate second-order random field, or a set of m-variate real

- To see if $\hat{f}(x) = x'\beta$ is a good candidate, we examine estimator and prediction error (how close they are to reality): $\hat{\beta} - \beta$?, 2) $y_0 - \hat{f}(x)$? Good estimators should, on average have, small prediction errors. Prediction after variable selection is biased (my paper: because β_{OLS} is unbiased!). Bias $\hat{\beta}$ through discrete variable selection or regularization. Econ fret about bias. Bias just one type of error. *Estimator* bias wrt truth can come from functional form of estimator f or from the conditioning set. Mis-specification bias, etc. *Unbiased* estimation of causal parameters (in non-parametric SCM) requires knowing the DAG (not necessarily the functional forms. Are unbiased estimates identifiable (Pearl). Pearl doesn't discuss bias-variance. Bias from algorithm (penalized OLS) vs model (mis-specification, Pearl). Discrete and continuous regularization (e.g. variable selection vs penalty). Some emphasize bias, others error (bias + variance).
- "penalty parameter", "ridge/lasso parameter", "hyper parameter". Higher ridge hparams bias penalized maximum likelihood estimates towards zero (shrinkage). Ridge L2 gives an isometric bias (rotatable, iid priors), cf lasso?

Technical challenges

- Model evaluation
- Relative: comparison/test (over-fitting spatial structure), e.g. permutation test p. 132
- Absolute: posterior predictive check, cross-validation, ...
- $[Y(\cdot), \theta | Z]$
- Fit/posterior parameters of covariance function parameters p.132
- Code
- - Spatial Rasch Models (package:inla, ...). A multilevel Rasch model can be estimated using the package lme4, nlme, and MCMCglmm with functions for mixed-effects models with crossed or partially crossed random effects. The ordinal package implements this approach for polytomous models. An infrastructure for estimating tree-structured item response models of the GLMM family using lme4 is provided in irtrees.
- - ...

Spatial LD/DIF

"Clustering", i.e. grouping, a proxy to local dependence (in the marginal model) or "coarse confounding/common causes" (i.e. school explains intelligence change).
"In vivo animal experiments demonstrate neurotoxicity of exposures to particulate matter (PM) and ozone, but only one small epidemiological study had linked ambient air pollution with central nervous system (CNS) functions in children" (Chen, Jiu-Chiuan, and Joel Schwartz. "Neurobehavioral effects of ambient air pollution on cognitive performance in US adults." Neurotoxicology 30.2 (2009): 231-239.)
Geostatistical methods in order to obtain a respiratory health risk map.
Benefits of population modeling for
Individual level: Optimal spatial prediction (kridging) of ability, e.g. to optimize test, e.g. by select items close to subjects' latent ability.
Population level: ...

Spatial LD/DIF

- Differential item functioning (DIF) can lead to an unfair advantage or disadvantage for certain subgroups.
- DIF: outcome probability varies (with some pre-specified x discrete variable, e.g. race, or unsupervised - latent class mixture models) DESPITE conditioning on latent variables. $p(y|x, \theta) \neq p(y|x', \theta)$, JC. Variation in success is ARTIFACT of test. Parameter "instability"/non-invariance means context-dependent parameters, i.e. interaction.
- spatio-temporal - c.f. psychological - sources for the differential functioning of the items.
- Rost (1990) mixture models considered as the most stringent test for the Rasch model, because it tests for item parameter differences between all possible groups of subjects regardless of person covariates. of these methods are designed for the comparison of pre-specified focal and reference groups, such as males and females. Latent class approaches, on the other hand, allow to detect previously unknown groups exhibiting DIF. However, this approach provides no straightforward

Conditional independence

Conditional independence is the unifying concept of importance for item response models and chain-graph models. We write XY—Z to indicate that two sets of variables, X = (X1,. . . ,Xa) and Y = (Y1,. . . ,Yb), are conditionally independent given a third set, Z = (Z1,. . . ,Zc), in the sense that P(X—Y,Z) = P(X—Z).

This chapter views Rasch analysis as an examination of the items given the requirements of ideal measurements, yielding a summary of problems and an evaluation of their relevance.
DIF (Rosenbaum, 1989) constant relation between latent traits and the items (no interactions): implies critereon validity, necessary for construct validity. DIF definition was vague. . We assume that it refers to meaningful and relevant partitions of the persons defined by an exogenous variable, but notice that in most studies a limited number of such variables will be available.
Absence of DIF can be stated as the requirement, Y X — , of conditional independence and because local independence implies pairwise conditional independence criterion-related construct validity defines a chain-graph model.
Log-linear models (for natural language processing, for contingency table modeling, for spatial modeling (ising models on a graph or metric distance features). Loglinear models for classification/count data. Distribution over tags in a sentence.
Regularization leads to better generalization, in cases with perhaps large numbers of features. log-linear, quite flexible in terms of the

types of features included. In supervised learning, features, eg binary features, $f_k(x, y)$. equivalence between (appropriately reparameterized) poisson and multinomial models, of which the bernoulli version is a special case. The equivalence you might be thinking where some variables are treated as explanatory and therefore conditioned on in the multinomial / logit analysis but treated as random and therefore modelled in the poisson / log-linear analysis.

Lfffffocal dependence

The physical functioning (PF) subscale summarizes responses to ten items under the common heading Does your health now limit you in these activities? If so, how much? Vigorous activities, e.g., running, heavy lifting, strenuous sport (PF1) Moderate activities (PF2) Lifting or carrying groceries (PF3) Climbing several fligths of stairs (PF4) Climbing one flight of stairs (PF5) Bending, kneeling, or stooping (PF6) Walking more than a mile (PF7) Walking several blocks (PF8) Walking one block (PF9) Bathing or dressing yourself (PF10) Three ordinal response categories (Not limited, Limited a little, Limited a lot) are used. The developers claim that Studies to date have yielded content, concurrent, criterion, construct, and predictive evidence of validity. Scrutinizing the items will show that LD between PF4 and PF5 and between PF7, PF8, and PF9 must be expected if responses are rational and consistent. Whether the reported analyses of construct validity may have overlooked this is not the focus of the present chapter. Problems of this kind are not unusual in

Technical challenges

- Stationarity
- f-valued parameters: mean function, covariance function, expected difference function (variagram), ...
- Optimal parameter estimation. covariance function (kernels) ¡- -¿ power spectral functions.
- Non-stationary variogram $2\gamma_Y(u,v) = var(Y(u) - Y(v)) = C_Y(u,v) + C_Y(u,v) - 2C_Y(u,v)$
- Non-stationary variogram $2\gamma_Y(s_1,s_2) = var(Y(s_1) - Y(s_2)) = C_Y(s_1,s_1) + C_Y(s_2,s_2) - 2C_Y(s_1,s_2)$, where s_i = subject in space i.
- fit covariance/spectral function
- optimal spatially prediction (kridging of geology/hydrology/ecology).
- Predicting uncertain $\{Y(s_{new}) : s_{new} \in D_s)\}$.
- Can view kridging as an HM p. 136

Technical challenges

- Gaussians closed under sum and product rule (marginalization/projection, conditioning), new parameters are linear algebraic functions of old parameters.
- Gaussian prior and linear function (for fixed covariates x) gives Gaussian observations.
- Not because its right, but because its tractable
- choose the kernel function. combine kernels algebraically
- Bayesian maximum a posteriori (posterior predictive mean) is frequentist L_2-regularized least squares loss (kernel ridge regression), i.e. over-fitting regularizers are "priors" (in the sense they regularize the location of maximum)
- infer hyperparameters: hierarchical bayesian inference or cross-validation
- GP is a distribution over a (continuous) set of real numbers that covary
- Euclidian "embeddings" of non-Euclidian - e.g. structured - objects such as fruit
- kridging, ridging, posterior predictive GP mean of latent